AF 1631

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

Appl. No.	:	09/921,045	Confirmation No.: 7635
Applicant	:	David Dorris, et al.	
Filed	:	August 2, 2001	
TC/A.U.	:	1631	
Examiner	:	Cheyne D. Ly	
Docket No.	:	PU01111	
Customer No.	:	22840	

Mail Stop Appeal Brief – Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

September 15, 2005

AMENDED APPEAL BRIEF

Sir:

In response to the Notification of Non-Compliance having the mailing date of September 12, 2005, Appellants submit this third amended Appeal Brief in triplicate, appealing from the June 14, 2004, rejection of the Primary Examiner, finally rejecting claims 1–3, 5, 16–18, 20 and 23-50 in the captioned application. The Notice of Appeal was filed on September 17, 2004, which contained authorization to charge the "Appeal Fee" to Appellants' Deposit Account. The Appeal Brief filed on November 15, 2004, was filed with a Transmittal of Appeal Brief (Large Entity), in duplicate, which contained authorization to charge the fee for filing the Appeal Brief to Appellants' Deposit Account. Appellants do not believe any additional fees are required. However, Appellants' hereby authorize the Commissioner to debit any fees due and credit any overcharges to Appellants' Deposit Account No. 502-590.

Real Party in Interest

Amersham Biosciences AB, formerly known as Amersham Pharmacia Biotech AB, the assignee and owner of the captioned application, is the real party in interest to this appeal.

Related Appeals and Interferences

There are no other appeals or interferences related to the instant appeal.

Status of Claims

Claims 1-67 are pending in the captioned application. Claims 4, 6-15, 19, 21, 22, and 51-67 have been withdrawn from consideration. The claims currently under examination, namely claims 1-3, 5, 16-18, 20 and 23-50 are appended hereto.

Status of Amendments

There are no outstanding amendments with regard to the captioned application.

Summary of Claimed Subject Matter

This invention provides methods of choosing probes to a target sequence (e.g., a gene), particularly probes for use in high-density oligonucleotide arrays (page 5, lines 21–22). Because the present method provides for probes that allow for the accurate determination of the amount of target sequence within a composition over a wide range of concentrations, the use of multiple probes per gene may be obviated (page 5, lines 23–25). Thus, such probes are useful in methods of accurately analyzing the expression of a

gene within a cell or group of cells using only a single probe (page 5, lines 25–28). The present invention also provides oligonucleotide arrays comprising such probes that are useful for accurately analyzing, at the same time, the expression (page 5, line 28 to page 6, line 1). The claims are directed to embodiments of the methodology.

Independent claim 1 recites a method wherein "three or more candidate probes are hybridized with a first composition comprising the target nucleic acid. A first hybridization signal is then determined for each of the candidate probes. (The determination of the hybridization signal may be repeated several times for each candidate probe and an average of all the determinations may be used in subsequent steps of the method). The candidate probes are then hybridized with a second composition comprising the target nucleic acid and a second hybridization signal is determined. A hybridization signal ratio is then calculated for each candidate probe. This ratio is the ratio of the first hybridization signal to the second hybridization signal for each candidate probe. These hybridization signal ratios from all the candidate probes are then averaged. The hybridization signal ratio from each candidate probe is then compared to the average hybridization signal ratio in order to chose which of the candidate probes is the appropriate probe for that target sequence. In a preferred embodiment, the candidate probe having a hybridization signal ratio closest to the average hybridization signal ratio is chosen".

Grounds of Rejection to be Reviewed on Appeal

1. Whether claims 17, 18, 20, and 23-28 are properly rejected under 35 U.S.C. § 112, second paragraph.
2. Whether claims1-3, 5, 16-18, 20 and 23-50 are properly rejected under 35 U.S.C. § 103(a) as being unpatentable over Manduchi et al. (2000) taken with Allzadeh et al. (2000) in combination with Lockhart et al. (US Pat. No. 6,040,138).

All of the rejected claims in the rejection appealed hereunder stand or fall together.

Argument

The Examiner has rejected claims 17, 18, 20 and 23–28 under 35 U.S.C. § 112, second paragraph as "being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention".

Specifically, the Examiner stated, "specific to claims 17 and 23-28, line 2, the term 'complementary' causes the claim to be vague and indefinite because it is not clear what criteria are being used to determine that a nucleic acid sequence is complementary to another. Is a complement of 2 nucleotides of two different nucleotides sequence [sic.] sufficient to consider said sequence complementary? Clarification of the metes and bounds is required. Claims 18 and 20 are rejected for being dependent from claim 17".

In response, Appellants respectfully asserted that the term "complimentary" is clear to one skilled in the art, and specifically directed the Examiner's attention to page 4, lines 4–11, wherein "complimentary" sequences are discussed. Inasmuch as claim 17 is dependent upon claim 1, which is directed to "a method of selecting a probe for a target nucleic acid sequence", Appellants respectfully asserted that requiring further metes and bounds is improper, as one skilled in the art could easily determine which probes are complimentary and which are not, and no metes and bounds are specifiable.

In response the Examiner stated that Appellants' arguments have "been fully considered and found to be unpersuasive", noting that the "pointed to support provides disclosure of one or more types of chemical bonds and recited 'Watson Crick base pairing' as an example" but continuing "said support does not provide criteria to which one of ordinary skill in the art may apply for determining a nucleic acid sequence being complementary to another sequence as specified by the elected claims".

In response, Appellants reiterate the arguments presented above and respectfully assert that one skilled in the art would readily comprehend what is meant by a complementary sequence, inasmuch as complementarity and pairing of bases (or which Watson Crick base pairing is a well recognized example) and/or modified bases is well recognized in the art. Appellants respectfully submit that the skill of the art is fairly advanced in this area, and that the skilled artisan would readily recognized what is and is not encompassed by the language the Examiner finds objectionable.

In view of the foregoing, Applicants respectfully assert the Examiner's rejections cannot be sustained and should be withdrawn.

The Examiner has rejected claims 1–3,5, 16–18, 20, 23–50 under 35 U.S.C. § 103(a) as "being unpatentable over Manduchi et al. (2000) taken with Allzadeh et al. (2000) in combination with Lockhart et al. (US 6,040,138 A)".

Specifically, the Examiner stated, "Manduchi et al. discloses a method for selection of a probe on a microarray for a target nucleic acid sequence wherein two samples types (first composition and second composition) and the ratios from the two separate two-channel microarrays are compared using the same reference for one of the channels...For each homotypic group and for each gene tag, Manduchi et al. computes the average intensity of that tag over a plurality of samples in the group, sets the group in order, establishes a reference group to which other groups are compared, and lists the ratios..."

The Examiner continued, "up regulation is determined by comparing ratio r^1, of the average intensity of a gene tag at group I and the average intensity of the same gene tag at the reference group...The method of Manduchi et al. is applied to hybridizing 3 or more candidate probes...generating datasets containing five homotypic groups comprising human blood progenitor cells..., as in instant claims 1, 3, 5, 20, 29, 31, and 40-50".

The Examiner further continued, "it is noted Manduchi et al. discloses a method directed to highly parallel gene expression experiments,...Although Manduchi et al. demonstrates said method with data generated from a two-channel microarray, said method is applicable to many types of data generated from highly parallel hybridization array experiments. It is well know in the art that a type of highly parallel hybridization array experiment is oligonucleotide arrays wherein gene expression is detected by the complementary of probe sequence to target sequence. The inclusion of a reference by Lipshutz et al. is not being used as prior art but to expand on what is well known in the art of parallel hybridization array experiments...Further, the inclusion of the Duggan et al. reference is not being used as prior art but to expand on what is well known in the art of parallel hybridization array experiments..."

The Examiner conceded, "Manduchi et al....does not disclose the limitations of claims 2, 30, and 32-37". However, the Examiner stated, "Allzadeh et al. discloses a method of generating said data by hybridizing select gene probes on a 'lymphochip' (first partner) to labeled targets from a cDNA libraries (second partner comprises a label)..." Further, the Examiner states, "the samples for microarray analysis disclosed by Allzadeh et al. comprises a low and high concentration and samples are treated in such growth conditions as phorbol ester, ionomycin, or anthracycline..."

The Examiner further conceded, "Manduchi et al. and Allzadeh et al. do not disclose the limitation wherein the first or second binding partner comprises biotin". The Examiner continued, "Lockhart et al. discloses the use of labels such as biotin for nucleic

acids (probe or target) in expression monitoring by hybridization to high-density oligonucleotide arrays...Lockhart et al. suggested an improvement for monitoring gene expression via hybridization arrays by using a rapid and effective method for identifying a set of oligonucleotide probes that maximized specific hybridization efficacy....The improvement suggested by Lockhart et al. is directly applicable to the method of parallel gene expression experiments via hybridization arrays..."

The Examiner concluded, "an artisan of ordinary skill in the art at the time of the instant invention would have been motivated by the improvement suggested by Lockhart et al. to perform a method of parallel gene expression experiments via hybridization arrays as taught by Manduchi et al. and Allzadeh et al. using biotin as taught by Lockhart et al. Therefore, it would have been obvious to one having ordinary skill in the art at the time of the invention was made to perform method of parallel gene expression experiments via hybridization arrays with biotin as taught by Manduchi et al., Allzadeh et al., and Lockhart et al.".

In response, Appellants respectfully submitted that the Examiner had misapplied the teachings of the base reference, Manduchi, et al., to the instant claimed invention. Specifically, as recited in claim 1, the invention encompasses a method for selecting a probe for a target nucleic acid molecule comprising seven steps. These steps include "hybridizing three or more candidate probes with a first composition comprising the target nucleic acid sequence; determining a first hybridization signal for each candidate probe; hybridizing the three or more candidate probes with a second composition

comprising the target nucleic acid sequence; determining a second hybridization signal for each candidate probe; calculating a hybridization signal ratio of the first hybridization signal to the second hybridization signal for each candidate probe; calculating an average hybridization signal ratio for the three or more candidate probes; and selecting the candidate probe by comparing a candidate probe's hybridization signal ratio to the average hybridization signal ratio".

Thus, Appellants summarized, the instant invention provides a process for identifying and selecting the best probe for a specific target nucleic acid sequence.

The Appellants further asserted that the Manduchi, et al. paper, on the other hand, discloses (in the abstract) a "protocol...to attach expression patterns to genes represented in a collection of hybridization array experiments", which method "reflects the broader change of focus in the field from studying a few genes with many replicates to studying many (possible thousands) of genes simultaneously, but with relatively few replicates. The approach of the instant differs from standard methods in that it exploits the fact that there are many genes in the arrays. These are used to estimate for each sample type an appropriate distribution that is employed to control false positives for each of the predictions made. Satisfactory results can be obtained using this method with as few as two replicates".

Thus, Appellants submitted, the Manduchi, et al. methodology exploits the fact that many assay determinations are being made in parallel, and such determinations may

be made with many genes. Indeed, at page 686, column 2, the authors state, "our method exploits the fact that there are hundreds of genes to estimate the appropriate gene independent distribution within each sample type. By integrating over these distributions, false positive rates are calculated directly". The Appellants concluded that the Manduchi, et al. reference discloses a methodology whereby the distribution of particular genetic markers, which may number many, can be determined among various cell types, further noting that there is no disclosure nor even any suggestion of a methodology to determine appropriate probes on the basis of the probes' hybridization signal ratio to the average hybridization signal ration of other probes for the same nucleic acid sequence.

Appellants further asserted that the addition of the Allzadeh article and the Lockhart, et al. patent do nothing to remedy this deficiency, specifically noting that the Lockhart, et al. patent is mentioned in the Background of Invention section of the captioned application at page, lines 15–22, wherein it is stated, "Lockhart *et al.*, ...describe...a method...in that...a number of candidate probes to a target sequence are tested to determine which probe provided the strongest signal". In an attempt to account for probes that show a high background signal even in the absence of the target, Lockhart *et al.* compare the probe signal to a signal obtained from a second probe constructed to contain a single mismatch with the target sequence. Only those probes having a signal that is a certain percentage over the signal obtained with the mismatch probe are used. Lockhart *et al.* describe using multiple probes for a given target sequence in an array to accurately determine the expression level of a gene over a range of concentrations".

However, as stated in the captioned application at lines 25, et. seq., "ideally an array would contain only one probe of each gene yet would still be able to provide accurate differential gene expression profiles. Because a probe giving the highest hybridization signal in the given concentration of intended target (chosen by rapid prototyping) [such as in Lockhart, et al.] may not always provide for accurate gene expression profiles wherein different samples having varying amounts or varying structures of the intended target, there is a need for arrays containing only a single probe to each gene yet are still able to indicate variation in the expression level of the gene".

Regarding the Allzadeh, et al. reference, Appellants asserted that there is similarly no disclosure or even any suggestion of the instant invention. Indeed, while Applicants concede that the Allzadeh, et al. article discloses gene expression profiling, such is quite different from the instant invention.

In response, the Examiner stated, "Claim 1 recites steps a) to g) for selecting a probe for a target nucleic acid; however, said claim does not recite [sic.] any limitations or methods or steps for determining the "best" or "appropriate probes". While Appellants do concede that these words are not used, Appellants also point out that the purpose of running multiple candidate probes in the claim is to determine which probe produces the desirable signal as compared with the average signal (see e.g. g) of Claim 1); indeed, there would be little reason to run the multiple if one were not comparing their performance of them. And this is neither disclosed, nor even suggested, by the combination of references set forth by the Examiner.

In response, the Examiner noted that Manduchi et al. discloses a method "wherein two sample types (first composition and second composition) and the ratios from the two separate-two channel microarrays are compared using the same reference for the channels ... For each homotypic group and each gene tag Manduchi et al. computes the average intensity of the tag over a plurality of samples in the group, sets the groups in order, establishes a reference group to which the other groups are compared, and lists the ratios". The Examiner continued, "Up regulation is determined by comparing the ratio of the average intensity of a gene tag at group I and the average intensity of the same tag at the reference group ... The reiterated dictation of Maduchi et al above is consistent with the limitations of steps a) to g) of claim 1".

The Examiner further states, "Specific to the argument that Manduchi et al. does not disclose or suggest a method for determining 'the best' or 'appropriate' probes ... said claim does not recites [sic.] any [such] limitations or steps".

In response, Appellants reiterate the argument presented above and specifically point out that step g) of claim 1 specifically requires selection of the ultimate probe for the nucleic acid sequence, "by comparing a candidate probe's hybridization signal to the average hybridization signal". Appellants respectfully assert that it is implicit in the claim that the probe selected by the methodology will be the best, or the appropriate probe for the nucleic acid sequence; otherwise, Appellants respectfully submit, the process would have no useful purpose.

Further, with regard to the Allzadeh and Lockhart et al. references, the Examiner states that the "combination of Manduchi et al., Allzadeh et al. has been directed to the limitations [of] claims 2, 30, and 32-37". The Examiner further cites a reference to Duggan et al. (Nature Genetics Supplement, 21, pp 10-14, January 1999) "to expand on what is known in the art of parallel hybridization" and specifically, the use of "cDNA microarrays wherein fluorescent tagged transcripts are, on average, 600 bp, have an average of 2 fluorescent tags per 100 bp and hybridize, all of the (contiguous) to their probe, and also cites a reference to Lipschutz et al. (Nature Genetics Supplement, 21, pp 20-24, January 1999) "to expand on what is known in the art of parallel hybridization" and further "wherein [a complementary] sequence is complementary to at least 15 contiguous nucleotides in the target sequence". The Examiner specifically states that the inclusion of each of these references "is not being used as prior art".

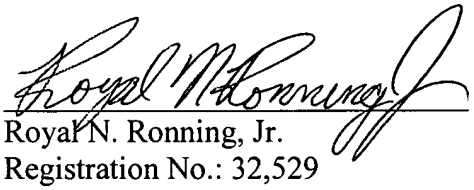
In response, Appellants reiterate the arguments presented above and respectfully assert that the inclusion of the Allazadeh and Lockhart et al. references, as well as the Duggan et al. and Lipschutz et al references ("not being used as prior art") do nothing to remedy the deficiencies of the Manduchi et al. reference and that the references, taken alone and in combination with one another, neither disclose nor even suggest the instant invention.

In view of the foregoing, Appellants respectfully submit that the Examiner's rejection cannot be upheld and should be reversed.

Conclusion

In view of the foregoing arguments, Appellants respectfully assert that the Examiner's rejections presented above cannot be sustained, and should be reversed.

Respectfully submitted,



Royal N. Ronning, Jr.
Registration No.: 32,529
Attorney for Appellants

Amersham Biosciences Corp
800 Centennial Avenue
P. O. Box 1327
Piscataway, New Jersey 08855-1327

Tel: (732) 457-8423
Fax: (732) 457-8463

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Mail Stop Appeal Brief – Patents, Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450, on September 15, 2005.

Signature: 

Name: Melissa Leck

CLAIMS APPENDIX

The Rejected Claims

Claim 1 (original): A method of selecting a probe for a target nucleic acid sequence, the method comprising the steps of:

a) hybridizing three or more candidate probes with a first composition comprising the target nucleic acid sequence;

b) determining a first hybridization signal for each candidate probe;

c) hybridizing the three or more candidate probes with a second composition comprising the target nucleic acid sequence;

d) determining a second hybridization signal for each candidate probe;

e) calculating a hybridization signal ratio of the first hybridization signal to the second hybridization signal for each candidate probe;

f) calculating an average hybridization signal ratio for the three or more candidate probes; and

g) selecting the candidate probe by comparing a candidate probe's hybridization signal ratio to the average hybridization signal ratio.

Claim 2 (original): The method of claim 1, wherein the target nucleic acid comprises cDNA.

Claim 3 (original): The method of claim 2, wherein the cDNA is derived from a mammalian cell.

Claim 4 (withdrawn): The method of claim 3, wherein the mammalian cell is a rat cell.

Claim 5 (original): The method of claim 3, wherein the mammalian cell is a human cell.

Claim 6 (withdrawn): The method of claim 1, wherein the target nucleic acid comprises genomic DNA.

Claim 7 (withdrawn): The method of claim 6, wherein the genomic DNA is derived from a mammalian cell.

Claim 8 (withdrawn): The method of claim 7, wherein the mammalian cell is a rat cell.

Claim 9 (withdrawn): The method of claim 7, wherein the mammalian cell is a human cell.

Claim 10 (withdrawn): The method of claim 1, wherein the target nucleic acid comprises RNA.

Claim 11 (withdrawn): The method of claim 10, wherein the RNA is derived from a mammalian cell.

Claim 12 (withdrawn): The method of claim 11, wherein the mammalian cell is a rat cell.

Claim 13 (withdrawn): The method of claim 11, wherein the mammalian cell is a human cell.

Claim 14 (withdrawn): The method of claim 1, wherein the target nucleic acid is derived from a prokaryote.

Claim 15 (withdrawn): The method of claim 1, wherein the target nucleic acid is derived from a virus.

Claim 16 (original): The method of claim 1, wherein the three or more candidate probes comprise a nucleic acid sequence complementary to the target sequence.

Claim 17 (original): The method of claim 1, wherein the three or more candidate probes comprise a nucleic acid sequence complementary to an expressed sequence or the expressed sequence's complement.

Claim 18 (original): The method of claim 17, wherein the expressed sequence comprises a mammalian expressed sequence.

Claim 19 (withdrawn): The method of claim 18, wherein the mammalian expressed sequence is a rat expressed sequence.

Claim 20 (original): The method of claim 18, wherein the mammalian expressed sequence is a human expressed sequence.

Claim 21 (withdrawn): The method of claim 1, wherein the three or more candidate probes comprise a nucleic acid sequence complementary to a genomic nucleic acid sequence.

Claim 22 (withdrawn): The method of claim 1, wherein the three or more candidate probes comprise a nucleic acid sequence complementary to a viral nucleic acid sequence or the viral nucleic acid sequence's complement.

Claim 23 (original): The method of claim 1, wherein the three or more candidate probes comprise a candidate probe comprising a nucleic acid sequence complementary to at least 15 contiguous nucleotides of the target sequence.

Claim 24 (original): The method of claim 23, wherein each of the three or more candidate probes comprise a nucleic acid sequence complementary to at least 15 contiguous nucleotides of the target sequence.

Claim 25 (original): The method of claim 1, wherein the three or more candidate probes comprise a candidate probe comprising a nucleic acid sequence complementary to at least 30 contiguous nucleotides of the target sequence.

Claim 26 (original): The method of claim 25, wherein each of the three or more candidate probes comprise a nucleic acid sequence complementary to at least 30 contiguous nucleotides of the target sequence.

Claim 27 (original): The method of claim 23, wherein the three or more candidate probes comprise a candidate probe comprising a nucleic acid sequence complementary to less than 100 contiguous nucleotides of the target sequence.

Claim 28 (original): The method of claim 25, wherein the three or more candidate probes comprise a candidate probe comprising a nucleic acid sequence complementary to less than 100 contiguous nucleotides of the target sequence.

Claim 29 (original): The method of claim 1, wherein a nucleic acid array comprises the three or more candidate probes.

Claim 30 (original): The method of claim 1, wherein the first composition and the second composition comprise a concentration of the target sequence, the concentration within the first composition differing from the concentration within the second composition.

Claim 31 (original): The method of claim 30, wherein the first composition is derived from a different tissue type from that in which the second composition is derived.

Claim 32 (original): The method of claim 30, wherein the first composition and the second composition are derived from a cell type grown at growth conditions, the growth conditions from which the first composition is derived differing from the growth conditions from which the second composition is derived.

Claim 33 (original): The method of claim 30, wherein the first composition and the second composition comprise different concentrations of a stock composition derived from one or more cells.

Claim 34 (original): The method of claim 1, wherein the hybridizing comprises stringent conditions.

Claim 35 (original): The method of claim 1, wherein the target nucleic acid comprises a detectable moiety.

Claim 36 (original): The method of claim 1, wherein the target nucleic acid comprises a first partner of a binding pair.

Claim 37 (original): The method of claim 36, wherein a second partner of the binding pair comprises a label.

Claim 38 (original): The method of claim 36, wherein the first partner comprises biotin.

Claim 39 (original): The method of claim 37, wherein the second partner comprises biotin.

Claim 40 (original): The method of claim 1, wherein determining a first hybridization signal comprises averaging more than one hybridization signal for the candidate probe hybridized with the first composition.

Claim 41 (original): The method of claim 1, wherein determining a second hybridization signal comprises averaging more than one hybridization signal for the candidate probe hybridized with the second composition.

Claim 42 (original): The method of claim 1, further comprising the steps of:

c1) hybridizing the three or more candidate probes with a third composition comprising the target nucleic acid sequence;

d1) determining a third hybridization signal for each candidate probe;

e1) calculating a second hybridization signal ratio of the first hybridization signal to the third hybridization signal for each candidate probe;

f1) calculating an average second hybridization signal ratio for the three or more candidate probes; and

g1) selecting the candidate probe by comparing a candidate probe's second hybridization signal ratio to the average second hybridization signal ratio

Claim 43 (original): The method of claim 42, wherein the selecting comprises selecting the candidate probe by comparing the candidate probe's hybridization signal ratio and second hybridization signal ratio to the average hybridization signal ratio and average second hybridization signal ratio.

Claim 44 (original): The method of claim 42, further comprising the steps of:

e2) calculating a third hybridization signal ratio of the second hybridization signal to the third hybridization signal for each candidate probe; and

f2) calculating an average third hybridization signal ratio for the three or more candidate probes.

Claim 45 (original): The method of claim 44, wherein the selecting comprises selecting the candidate probe by comparing the candidate probe's hybridization signal ratio, second hybridization signal ratio, and third hybridization signal ratio to the average hybridization signal ratio, average second hybridization signal ratio, and average third hybridization signal ratio.

Claim 46 (original): The method of claim 1, wherein selecting comprises selecting the candidate probe having a hybridization signal ratio closest to the average hybridization signal ratio.

Claim 47 (original): The method of claim 42, wherein selecting comprises selecting the candidate probe having a second hybridization signal ratio closest to the average second hybridization signal ratio.

Claim 48 (original): The method of claim 43, wherein the selecting comprises selecting the candidate probe having a hybridization signal ratio and second hybridization signal ratio closest to the average hybridization signal ratio and average second hybridization signal ratio.

Claim 49 (original): The method of claim 45, wherein the selecting comprises selecting the candidate probe having a hybridization signal ratio, second hybridization signal ratio, and third hybridization signal ratio closest to the average hybridization signal ratio, average second hybridization signal ratio, and average third hybridization signal ratio.

Claim 50 (original): The method of claim 1, wherein the first composition comprises a first concentration of the target nucleic acid sequence and the second composition comprises a second concentration of the target nucleic acid sequence, the method comprising:

alternatively to step f), a step of calculating a concentration ratio of the first concentration of the target nucleic acid to the second concentration of the target nucleic acid; and

alternatively to step g), selecting the candidate probe by comparing the candidate probe's hybridization signal ratio to the concentration ratio.

Claim 51 (withdrawn): The method of claim 66, wherein the selecting comprises selecting the candidate probe having a hybridization signal ratio closest to the concentration ratio.

Claim 52 (withdrawn): A method of making an oligonucleotide array, comprising the steps of:

a) hybridizing three or more candidate probes comprising a nucleic acid sequence with a first composition comprising the target nucleic acid sequence;

b) determining a first hybridization signal for each candidate probe;

c) hybridizing the three or more candidate probes with a second composition comprising the target nucleic acid sequence;

d) determining a second hybridization signal for each candidate probe;

e) calculating a hybridization signal ratio of the first hybridization signal to the second hybridization signal for each candidate probe;

f) calculating an average hybridization signal ratio for the three or more candidate probes;

g) selecting the candidate probe by comparing the candidate probe's hybridization signal ratio to the average hybridization signal ratio, yielding a first probe; and

h) constructing an oligonucleotide array comprising a probe comprising the nucleic acid sequence of the first probe.

Claim 53 (withdrawn): The method of claim 52, wherein steps a) through g) are repeated with a second target sequence and second candidate probes to yield a second probe and constructing a nucleic acid array comprising the first probe and the second probe.

Claim 54 (withdrawn): The method of claim 52, wherein selecting comprises selecting the candidate probe having a hybridization signal ratio closest to the average hybridization signal ratio.

Claim 55 (withdrawn): An oligonucleotide array comprising at least 10 probes to 10 different human genes, the probes selected using the method of claim 1.

Claim 56 (withdrawn): The oligonucleotide array of claim 55 comprising at least 100 probes to 100 different human genes, the probes selected using the method of claim 1.

Claim 57 (withdrawn): The oligonucleotide array of claim 56 comprising at least 1000 probes to 1000 different human genes, the probes selected using the method of claim 1.

Claim 58 (withdrawn): The oligonucleotide array of claim 57 comprising at least 5000 probes to 5000 different human genes, the probes selected using the method of claim 1.

Claim 59 (withdrawn): The oligonucleotide array of claim 58 comprising at least 10000 probes to 10000 different human genes, the probes selected using the method of claim 1.

Claim 60 (withdrawn): The oligonucleotide array of claim 55, wherein every probe of the array represents a different gene.

Claim 61 (withdrawn): The oligonucleotide array of claim 58 wherein every probe of the array represents a different gene.

Claim 62 (withdrawn): A method of analyzing the expression of a gene within a source, comprising:

a) hybridizing a nucleic acid composition derived from the source with the oligonucleotide array of claim 47 comprising a probe representing the gene; and

b) determining hybridization of a nucleic acid within the composition to the probe representing the gene, wherein hybridization of a nucleic acid within the composition to the probe representing the gene indicates expression of the gene within the source.

Claim 63 (withdrawn): The method of claim 62, wherein the expression of at least 10 genes is analyzed.

Claim 64 (withdrawn): The method of claim 63, wherein the expression of at least 100 genes is analyzed.

Claim 65 (withdrawn): The method of claim 64, wherein the expression of at least 1000 genes is analyzed.

Claim 66 (withdrawn): The method of claim 65, wherein the expression of at least 5000 genes is analyzed.

Claim 67 (withdrawn): The method of claim 66, wherein the expression of at least 10000 genes is analyzed.

EVIDENCE APPENDIX

Applicants hereby append a copy of:

U.S. Patent 6,040,138 by Lockhart et al.;

Manduchi, E., et al., "Generation of patterns from gene expression data by assigning confidence to differentially expressed genes", *Bioinformatics*, volume 16, number 8, 2000, pages 685-698;

Allzadeh, A., et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", *Nature*, volume 403, 2000, pages 503-511.

These are the evidence relied upon by the Examiner for rejection of appealed claims.

BIOINFORMATICS

Vol. 16 no. 8 2000
Pages 685–698

Generation of patterns from gene expression data by assigning confidence to differentially expressed genes

Elisabetta Manduchi[1,*], Gregory R. Grant[1], Steven E. McKenzie[2], G. Christian Overton[1], Saul Surrey[2] and Christian J. Stoeckert Jr.[3]

[1]Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, USA, [2]Hematology/Oncology Research, A.I. duPont Hospital for Children, Wilmington, DE 19803, and Department of Pediatrics, Jefferson Medical College, Philadelphia, PA 19107, USA and [3]Division of Hematology, The Children's Hospital of Philadelphia and Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, USA

Received on December 15, 1999; revised on March 14, 2000; accepted on March 21, 2000

Abstract

Motivation: *A protocol is described to attach expression patterns to genes represented in a collection of hybridization array experiments. Discrete values are used to provide an easily interpretable description of differential expression. Binning cutoffs for each sample type are chosen automatically, depending on the desired false-positive rate for the predictions of differential expression. Confidence levels are derived for the statement that changes in observed levels represent true changes in expression. We have a novel method for calculating this confidence, which gives better results than the standard methods. Our method reflects the broader change of focus in the field from studying a few genes with many replicates to studying many (possibly thousands) of genes simultaneously, but with relatively few replicates. Our approach differs from standard methods in that it exploits the fact that there are many genes on the arrays. These are used to estimate for each sample type an appropriate distribution that is employed to control the false-positive rate of the predictions made. Satisfactory results can be obtained using this method with as few as two replicates.*

Results: *The method is illustrated through applications to macroarray and microarray datasets. The first is an erythroid development dataset that we have generated using nylon filter arrays. Clones for genes whose expression is known in these cells were assigned expression patterns which are in accordance with what was expected and which are not picked up by the standards methods. Moreover, genes differentially expressed between normal and leukemic cells were identified. These included genes whose expression was altered upon induction of the leukemic cells to differentiate. The second application is to the microarray data by Alizadeh* et al. *(2000). Our results are in accordance with their major findings and offer confidence measures for the predictions made. They also provide new insights for further analysis.*

Availability: *Software is available on request from the authors.*

Contact: *manduchi@pcbi.upenn.edu*

Introduction

The goal of this paper is to provide tools for the investigator to aid in the analysis of data collected from highly parallel gene expression experiments, such as hybridization array experiments. In particular we wanted to generate descriptive, yet dependable, expression patterns representing the differential expression of genes across cell types. Figure 1 illustrates how an excerpt from a typical 'raw' input might be transformed into an easily interpretable list of patterns.

Suppose we are comparing between two sample types, type A and type B (e.g. these could be two different filter arrays, or the two channels on a microarray, or the red-to-green ratios from two separate two-channel microarrays using the same reference for one of the channels). With highly parallel experiments we are presented with gene expression data from hundreds to thousands of genes simultaneously. We wish to identify those genes that are 'most likely' to be differentially expressed. Variation in

*To whom correspondence should be addressed.

(a)

gene tag	G_0 I	G_0 II	G_0 III	G_1 I	G_1 II	G_2 I	G_2 II	G_2 III	G_3 I	G_3 III
1	0.0111	0.0328	0.0151	0.0060	0.0236	0.0136	0.5610	0.8920	0.0639	0.2190
2	0.0050	0.0131	0.0061	0.0041	0.0361	0.0296	0.8830	0.7000	0.0199	0.1650
3	0.0629	0.2310	0.0131	0.2270	0.2120	0.0105	0.1100	0.0213	0.0117	0.0907
4	0.0250	0.0600	0.0264	0.1500	0.2660	0.0131	0.1860	0.0851	0.0172	0.0112

(b)

gene tag	G_1	G_2	G_3
1	0	7	2
2	0	8	1
3	2	-1	-1
4	3	1	0

Fig. 1. Excerpt from typical input data (normalized intensities) for four gene tags (a) and the patterns associated to these data by our program (b). Note there are three replicate experiments for sample types G_0 and G_2, and two replicate experiments for types G_1 and G_3. In (b) positive (resp. negative) integers represent up-regulation (resp. down-regulation) with respect to the reference group G_0.

the data (biological and/or experimental) can result in the observed level for a given gene in type B to be higher than in type A when in fact the gene is not truly up-regulated in type B. However, if C is a constant and if the intensity of a gene in type B is more than C times the intensity in type A, then if C is sufficiently large it will be unlikely that this observation is due just to variation. If C is too large, then some true cases of up-regulation will be missed unnecessarily. Therefore, the determination of an appropriate C should be based on explicit measures of confidence in order to effectively guide the predictions about differential expression. The measures of confidence should be based on the variability in the data with replicates used to gauge this variation.

Claverie in his survey paper (Claverie, 1999) points out that most published studies are 'quite elusive about measurement reproducibility and the confidence levels of the observed changes in expression are rarely assessed using standard methods.' He examines several studies that do provide replicate data and evaluates their thresholds with the standard methods, showing that rather high thresholds must be chosen when there are only two or three replicates, in order to get significance levels of 5%. Furthermore, the significance measure computed is the probability of predicting a gene is differentially regulated when it is actually not (which we call *the false-positive rate*). Because we are searching for a small set of genes from a large pool, one needs a very small false-positive rate to have a reasonable confidence in the predictions. For example, suppose that 50 of 1000 genes are up-regulated in type B, and our false-positive rate is 0.05, then there will be on average 47.5 genes falsely predicted to be up-regulated, nearly as many as are truly up-regulated. So the confidence in the prediction that a gene is up-regulated cannot be much higher than 50%. For many applications, this confidence level is much too low to be acceptable. To get the false-positive rate down and maintain reasonable cutoffs, standard methods require many replicates (see Claverie, 1999), on the order of ten or more. We have developed a method which gives reasonable cutoffs with as few as two replicates and which gives false-positive rates low enough to maintain good confidence levels. Our method exploits the fact that there are hundreds of genes to estimate appropriate gene-independent distributions in each sample type. By integrating over these distributions, false-positive rates are calculated directly. Finally, measures of confidence are then computed using Bayes theorem. Note however, this is not a Bayesian approach, we simply use Bayes theorem to reverse the conditionality clauses to turn the false-positive rate into a confidence measure.

We note that in Chen *et al.* (1997) a statistical analysis procedure is presented to determine differential expression for the special case of comparisons between the two channels of a single microarray. In contrast, our techniques are general and can be applied to many types of data. In Section **Results** we illustrate applications of the method to macroarray hematopoietic data that we have generated. We also illustrate an application to two-channel microarray data described in Alizadeh *et al.* (2000). (In the latter the comparisons are between red-to-green ratios from separate two-channel microarrays using the same reference for the green channel.) Software implementing the pattern generating algorithm has been developed and is available from the authors together with the documentation.

We end this introduction by defining some terminology that will be used throughout this paper. The gene expression data is taken from a collection of

samples. Here we use the generic term 'sample' to denote either an individual cell, or a cell type, or a tissue type, at a certain time point and under certain conditions. In hybridization arrays, data for one sample are in the form of intensities of spots on the array, where each spot corresponds to a gene. Arrays usually have *gene tags* on them, of some sort or another, for example clones or oligonucleotides, and often they have different tags for the same gene. The intensities of such tags might be merged into one datum. However, all occurrences are not always known, so to be precise we will often need to refer to expression levels and expression patterns of gene tags rather than of genes themselves.

Methods and Algorithm

In this section we describe our protocol for defining expression patterns. The input consists of normalized data, where the normalization procedure depends on the kind of experiments conducted and is therefore left to the user. In Sections **Application to an erythroid development nylon filter dataset** and **Application to a two-channel microarray lymphocyte dataset** we describe the normalization procedures used respectively for our nylon filter data and for the two-channel microarray data of Alizadeh *et al.* (2000). The input normalized intensities are subjected to preprocessing steps, which might include a shift of the intensities by a constant. These steps, together with the details on the binning procedure, are presented in Figure 2 and also discussed in Sections **Implementation** and **Results**. The reason for and the effect of the numerical shift are discussed below and in Sections **Comparison with standard statistical analysis methods** and **Discussion**.

In what follows, we will assume that the necessary preprocessing steps have taken place (including all appropriate normalizations and elimination of data with values too low to be distinguishable from background and including any numerical shift) and we describe our rationale in assigning a pattern to each gene tag under consideration. Unless otherwise stated, the expression 'intensity' refers to the processed intensity.

For each sample type, experiments should be replicated one or more times. Replicates allow the variability to be estimated. If no replicates are given for a sample type, default values are chosen, however, one can clearly make no meaningful estimate about confidence for those sample types with no replicates. We will use the expression *homotypic group of samples*, or *homotypic group* for short, to refer to a set of samples of the same type, possibly consisting of just one sample if no replicate experiments are performed on that sample type. In each gene tag's expression pattern there will be one symbol for each homotypic group. For each homotypic group and for each gene tag, we compute the average intensity of that tag over those samples in the group which have values for that tag. This will represent the intensity of that tag at that group.

The assignment of an expression pattern to each gene tag is done in two stages. In the first stage, we attach to each tag an ordered list of real numbers. In the second stage, we bin the numbers in this list, resulting in a pattern of integers. Binning levels are chosen to take into account the variability within each homotypic group and to ensure an expected false-positive rate of $s\%$ for our predictions about up- or down-regulation.

For the first stage, we start by fixing an ordering of the groups in our collection. If the collection is an ordered series (e.g. a time series), then an ordering is given to us *a priori*. There will be some reference group to which we compare our groups, i.e. with respect to which up- and down-regulation are measured. The reference group is chosen by the user according to the questions posed. Moreover, the user might want to compare expression levels to the median of the group intensities in the data, rather than to a particular group. Thus, to each tag we attach the ordered list of real numbers obtained by dividing each of its (non-reference) group intensities by its group intensity at the provided reference group or, respectively, by the median of its group intensities. We will refer to this list as the list of *ratios* attached to that tag. We denote by ℓ the length of this list (this means that we started with $\ell+1$ homotypic groups, if ratios were taken to a reference group, or with ℓ homotypic groups, if ratios were taken to the median).

For the second stage, for each (non-reference) group, we partition the range $[0, \infty)$ into disjoint subintervals, which we call *bins*. These bins depend on the group. Thus each group has its own set of bins and the number of bins used also depends on the group. Before showing exactly how to choose the bins we establish some notation and describe how the binning takes place. For each group, we number the bins from left to right using consecutive integers $-m, \ldots, 0, \ldots, n$, where the bin labeled 0 is the one containing the ratio 1 (so it represents the 'no change level'). We then attach to each gene tag the (ordered) list of integers, which we will call *levels*, obtained by looking, for each group, to which bin the ratio value for the tag at that group belongs. This list will represent the *expression pattern* of that tag. More precisely, suppose that for group i we have subdivided $[0, \infty)$ into $m_i + n_i + 1$ bins as $[0, \infty) = B_{i,-m_i} \cup B_{i,-m_i+1} \cup \cdots \cup B_{i,0} \cup \cdots \cup B_{i,n_i}$, where $B_{i,-m_i} = [0, a_{i,1})$, $B_{i,-m_i+1} = [a_{i,1}, a_{i,2})$, $\ldots$, $B_{i,n_i} = [a_{i,m_i+n_i}, \infty)$, for some real numbers $a_{i,1} < a_{i,2} < \cdots < a_{i,m_i+n_i}$. If the list of ratios obtained in the first stage for a certain gene tag is $(r_1, r_2, \ldots, r_\ell)$, then each r_i $(i = 1, 2, \ldots, \ell)$ belongs to exactly one of the bins $B_{i,j}$ $(j = -m_i, \ldots, n_i)$, say r_i belongs to B_{i,j_i}. The expression pattern associated to this tag is then $(j_1, j_2, \ldots, j_\ell)$.

1. If the user provided the list of minimum useful values, for each $i = start, \ldots, \ell$ and for each $k = 1, \ldots, t_i$, let $muv_{i,k}$ be the given minimum useful value for the k-th sample of the i-th group. Go to step 2.
If the user chose to provide d instead, let d be as given. If the user provided neither d nor the list of minimum useful values, let $d = 100$. For each $i = start, \ldots, \ell$ and for each $k = 1, \ldots, t_i$, let $muv_{i,k} = x_{h_0,i,k}$, where $x_{h_0,i,k}$ is such that $d\%$ of the $x_{h,i,k}$'s with $x_{h,i,k} \neq undef$ are greater than or equal to $x_{h_0,i,k}$.

2. For each $i = start, \ldots, \ell$, let $muv_i = \frac{\sum_{k=1}^{t_i} muv_{i,k}}{t_i}$. If $start = 1$, let $muv_0 = \frac{\sum_{i=1}^{\ell} muv_i}{\ell}$.

3. Let $\mathcal{U} = \{h : 1 \leq h \leq N$ and $\forall i$ with $start \leq i \leq \ell$ there exists k with $1 \leq k \leq t_i$ and $x_{h,i,k} \neq undef\}$. For each $h \in \mathcal{U}$ and for each $i \in \{start, \ldots, \ell\}$, let

$$\bar{x}_{h,i} = \frac{\sum_{x_{h,i,k} \neq undef} x_{h,i,k}}{\sum_{x_{h,i,k} \neq undef} 1}.$$

If $start = 1$, for each $h \in \mathcal{U}$ let $\bar{x}_{h,0}$ be the median of the $\bar{x}_{h,i}$'s over $i = 1, \ldots, \ell$.

4. Let $\mathcal{N} = \{h \in \mathcal{U}:$ for each i with $0 \leq i \leq \ell$, $\bar{x}_{h,i} \geq muv_i\}$.

5. Let $m = \min \frac{\sqrt{t_i}\bar{x}_{h,i}}{s_{h,i}}$ and $\varsigma = \max \frac{s_{h,i}}{\sqrt{t_i}}$ where $s_{h,i}$ is the sample standard deviation of the $x_{h,i,k}$ and the min and max are taken over all $i = start, \ldots, \ell$ and $h \in \mathcal{N}$ such that $x_{h,i,k} \neq undef$ for each $k = 1, \ldots t_i$. For each $h \in \mathcal{N}$, $i = 0, \ldots, \ell$, and $k = 1, \ldots, t_i$, let $x_{h,i,k} = x_{h,i,k} + shift$ and $\bar{x}_{h,i} = \bar{x}_{h,i} + shift$, where $shift = \max\{0, (7 - m)\varsigma\}$.

6. Let $\mathcal{G} = \{h \in \mathcal{N} : \bar{x}_{h,0} \neq 0\}$. For each $h \in \mathcal{G}$ and for each $i = 1, \ldots, \ell$, let $r_{h,i} = \frac{\bar{x}_{h,i}}{\bar{x}_{h,0}}$. Then the list of ratios associated to the h-th gene tag is $(r_{h,1}, r_{h,2}, \ldots, r_{h,\ell})$.

7. For each $i = 1, \ldots, \ell$, let min_i (resp. max_i) be the minimum (resp. maximum) of $r_{h,i}$ over all $h \in \mathcal{G}$.

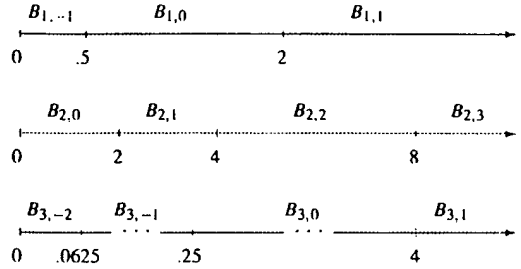
8. For each $i = 1, \ldots, \ell$, compute the upper cutratio C_i and the lower cutratio c_i as explained in section Methods and Algorithm.

9. For each $i = 1, \ldots, \ell$, the level cutoffs list is obtained by taking all successive powers of C_i which are strictly less then max_i and all successive powers of c_i which are strictly greater than min_i and for which there is at least one smaller non-zero $r_{h,i}$. This is the list $a_{i,1}, a_{i,2}, \ldots a_{i,s_i}$, with notation as in section Methods and Algorithm. Let $B_{i,-m_i}$, $B_{i,-m_i+1}$, $\ldots$, B_{i,n_i} denote respectively the intervals $[0, a_{i,1})$, $[a_{i,1}, a_{i,2})$, $\ldots$, $[a_{i,s_i}, \infty)$, where $B_{i,0}$ is the interval containing 1.

10. For each $h \in \mathcal{G}$ and for each $i = 1, \ldots, \ell$, let j_i be such that $r_{h,i} \in B_{i,j_i}$. The pattern attached to h is then $(j_1, j_2, \ldots, j_\ell)$.

Fig. 2. The pattern generation algorithm in version I of the software. N is the total number of distinct gene tags, $start = 0$, if a reference group is provided (in which case this is referred as 'group 0'), $start = 1$ otherwise, and the number of non-reference groups is ℓ. For each $i = start, \ldots, \ell$, the number of samples in the i-th group is denoted by t_i. For each $h = 1, \ldots, N$, $i = start, \ldots, \ell$, and $k = 1, \ldots, t_i$, $x_{h,i,k}$ is the (normalized) intensity of the h-th gene tag at the k-th sample in the i-th group, if this is given, otherwise $x_{h,i,k} = undef$. Finally, d is such that the top $d\%$ intensity levels are the ones likely to have given hybridization signals above background. Alternatively the user can specify, for each sample, a value ('minimum useful value') such that only intensities above this value are the ones likely to have given hybridization signals above background.

EXAMPLE. $\ell = 3$;
$m_1 = 1, n_1 = 1, B_{1,-1} = [0, 0.5), B_{1,0} = [0.5, 2), B_{1,1} = [2, \infty)$;
$m_2 = 0, n_2 = 3, B_{2,0} = [0, 2), B_{2,1} = [2, 4), B_{2,2} = [4, 8), B_{2,3} = [8, \infty)$;
$m_3 = 2, n_3 = 1, B_{3,-2} = [0, 0.0625), B_{3,-1} = [0.0625, 0.25), B_{3,0} = [0.25, 4), B_{3,1} = [4, \infty)$.



If the list of ratios for the gene tag in question is (0.2, 1.1, 4.1), the expression pattern attached to this tag is $(-1, 0, 1)$, since $0.2 \in B_{1,-1}$, $1.1 \in B_{2,0}$, and $4.1 \in B_{3,1}$.

We now explain how we choose the $a_{i,j}$'s (which we will refer to as *level cutoffs*). To illustrate the basic idea, suppose that we are taking ratios to a reference homotypic group, call it group 0 and let's focus on a fixed group i, $i \geq 1$. Suppose that we have replicate experiments for each of these two groups. We concentrate here on up-regulation; for down-regulation we proceed in an analogous fashion. Our goal is to achieve a certain degree of confidence in the assertion 'this gene is up-regulated at group i as compared to the reference group.' Each gene will have a certain (unknown) distribution of intensities in a group (reference or not), whose mean we will call 'the true mean intensity of the gene at that group.' Denote the random variable giving the intensity of gene g at group j by $X_{g,j}$ and denote the mean andstandard deviation of $X_{g,j}$ by $\mu_{g,j}$ and $\sigma_{g,j}$ respectively. By the statement 'gene g is up-regulated at group i as compared to the reference group' we mean that

$$\frac{\mu_{g,i}}{\mu_{g,0}} > 1.$$

We do not know these true means. We only know the observed intensities of the tags corresponding to g at the available replicates for each of the two groups. Fix such a tag h and denote the average of its observed intensities for group j by $\bar{x}_{h,j}$ and its observed intensity at the k-th replicate of this group by $x_{h,j,k}$. Let $s\%$ be the desired false-positive rate, that is the probability that we say that a gene tag h, corresponding to a gene g, shows up-regulation at group i as compared to the reference group, given that g is not up-regulated (i.e. given that $\frac{\mu_{g,i}}{\mu_{g,0}} \leq 1$). Our goal is to determine a value $C_i > 1$ (which we will call the *upper cutratio for group i*) such that, if we predict that a gene tag h is up-regulated at group i as compared to the reference group when $\frac{\bar{x}_{h,i}}{\bar{x}_{h,0}} > C_i$, then our false-positive rate is expected to be no greater than $s\%$. In order to do this, let $X_{g,j}$ be as above, and let $\bar{X}_{g,j} = (X_{g,j,1} + X_{g,j,2} + \cdots + X_{g,j,t_j})/t_j$, where t_j is the number of available replicates for group j and where the $X_{g,j,k}$'s are independent random variables each with the same distribution as $X_{g,j}$. Thus $\bar{x}_{h,j}$ is an observed value of $\bar{X}_{g,j}$ and $x_{h,j,k}$ an observed value of $X_{g,j,k}$. The false-positive rate is

$$\text{Prob}\left(\frac{\bar{X}_{g,i}}{\bar{X}_{g,0}} > C_i \,\middle|\, \frac{\mu_{g,i}}{\mu_{g,0}} \leq 1\right), \tag{1}$$

where the randomness is coming both from letting g vary over the genes and from the distributions of intensities for the genes. (1) is clearly less than or equal to

$$\text{Prob}\left(\frac{\frac{\bar{X}_{g,i}}{\mu_{g,i}}}{\frac{\bar{X}_{g,0}}{\mu_{g,0}}} > C_i \,\middle|\, \frac{\mu_{g,i}}{\mu_{g,0}} \leq 1\right). \tag{2}$$

Now, we claim that the events $\frac{\frac{\bar{X}_{g,i}}{\mu_{g,i}}}{\frac{\bar{X}_{g,0}}{\mu_{g,0}}} > C_i$ and $\frac{\mu_{g,i}}{\mu_{g,0}} \leq 1$ are independent. Note that the first inequality is the same as $\frac{\bar{X}_{g,i}}{\bar{X}_{g,0}} > C_i \frac{\mu_{g,i}}{\mu_{g,0}}$. Therefore this condition tells us that either $\mu_{g,i}$ is smaller than $\bar{X}_{g,i}$, or $\mu_{g,0}$ is larger than $\bar{X}_{g,0}$, or some combination of the two (how much smaller or larger depends on how large C_i is). In other words the condition is telling us how the sample means compare to the true means. This however should give no information on how the true means relate *to each other*. Therefore the events can reasonably be assumed to be independent. Thus the conditional probability above should equal the non-conditional probability

$$\text{Prob}\left(\frac{\frac{\bar{X}_{g,i}}{\mu_{g,i}}}{\frac{\bar{X}_{g,0}}{\mu_{g,0}}} > C_i\right). \tag{3}$$

So we are now seeking a $C_i > 1$ (as small as possible) such that

$$\text{Prob}\left(\frac{\frac{\bar{X}_{g,i}}{\mu_{g,i}}}{\frac{\bar{X}_{g,0}}{\mu_{g,0}}} > C_i\right) < s\%.$$

If we knew the distributions of $\frac{\bar{X}_{g,i}}{\mu_{g,i}}$ and of $\frac{\bar{X}_{g,0}}{\mu_{g,0}}$ (for g varying over the genes) we could compute such a C_i. But, the $\mu_{g,j}$'s are unknown. We approximate the distribution of $\frac{\bar{X}_{g,j}}{\mu_{g,j}}$ $(j = 0, i)$ for g varying over the genes by the distribution of

$$\frac{\frac{X_{g,j,k}}{\bar{X}_{g,j}} - 1}{\sqrt{t_j - 1}} + 1 \tag{4}$$

for g varying over the genes and k varying over the replicates for group j. In the Appendix we give a justification of this fact, in the case in which the intensity distribution of a gene in a group is close to normal and the distribution of ratios $(\sqrt{t_j})\mu_{g,j}/\sigma_{g,j}$, for g varying over the genes, is concentrated away from zero. The latter condition can be obtained by shifting the intensities appropriately so that the mean is increased while the standard deviation remains the same (see algorithm step 5 in Figure 2). The effect of this shift is discussed in Sections **Comparison with standard statistical analysis**

methods and **Discussion**. As we have investigated in simulations, approximation (4) holds for a much wider class of distributions than just normals (furthermore it improves quickly as t_j, the number of replicates, increases). In all cases of distributions which were not highly asymmetric, approximation (4) held well, and in most of the remaining cases it gave a conservative estimate, in that the approximation tended to have greater dispersion than the actual distribution. This failed most notably when the actual distribution was taken to be exponential, however this was not a relevant case to our applications, nor does it seem likely that it will arise in general.

Using this, we estimate the distribution of (4) from our data, i.e. from the distribution of

$$\frac{\frac{x_{h,j,k}}{\bar{x}_{h,j}} - 1}{\sqrt{t_j - 1}} + 1, \tag{5}$$

for h varying over all gene tags and k varying over the replicates for group j. Of course the more the gene tags and replicates, the better the latter approximation of (4) with an empirical distribution will be. Having a number of observations (number of gene tags $\times$ number of samples in group j) of at least 100 is suggested before adopting such an approximation. We then compute the desired C_i through integration. In particular, if f_j $(j = 0, i)$ is the density function for $\frac{\bar{X}_{g,j}}{\mu_{g,j}}$, and C is fixed, then we evaluate numerically (using the distribution of (5))

$$\int_t \int_{s>Ct} f_0(t) f_i(s) ds dt.$$

If this is above (resp. below) the desired false-positive rate, then C is raised (resp. lowered), and the integral is recalculated. C continues to be adjusted in this way (as a binary search) until the desired false-positive rate is attained. Then C_i is set to the value of C that gives this desired rate.

For down-regulation we proceed in an analogous manner and look for a $c_i < 1$ (as large as possible), which we call the *lower cutratio for group i*, such that

$$\text{Prob}\left(\frac{\frac{\bar{X}_{g,i}}{\mu_{g,i}}}{\frac{\bar{X}_{g,0}}{\mu_{g,0}}} < c_i \,\middle|\, \frac{\mu_{g,i}}{\mu_{g,0}} \geq 1 \right) < s\%.$$

To do this we use approximation (4) again.

Once the C_i's and c_i's have been determined for each non-reference group i, if the ratio r_i of the average intensity of a gene tag at group i and the average intensity of the same gene tag at the reference group (resp. the median) is between C_i and C_i^2 we say that the gene tag is up-regulated one level at this group as compared to the reference group (resp. the median). If r_i is between C_i^2 and C_i^3 then we say that the gene tag is up-regulated two levels as compared to the reference group, etc. A similar approach is taken for down-regulation. Thus we set the $a_{i,j}$'s to be

$$\ldots, c_i^2, c_i, C_i, C_i^2, \ldots.$$

Powers of C_i and c_i are used in order to respect proportions. The algorithm in Figure 2 (step 9) contains the details about how many of these powers we consider for each i.

Note that by having the false-positive rate, one can get a measure of the confidence in the statement that a gene is up-regulated at a group as compared to the reference group. Namely one can estimate the probability Prob(not up | predicted up) that a gene is not up-regulated given that we predict it is, by:

$$\begin{aligned} &\text{Prob(not up} \mid \text{predicted up)} \\ &\quad = \frac{\text{Prob(not up)}}{\text{Prob(predicted up)}} \text{Prob(predicted up} \mid \text{not up)} \\ &\quad \leq \frac{\text{Prob(predicted up} \mid \text{not up)}}{\text{Prob(predicted up)}}. \end{aligned}$$

where Prob(predicted up | not up) is the false-positive rate and Prob(predicted up) can be estimated empirically from the data, simply from counting the number of gene tags to which a positive level has been assigned by our protocol for the group under consideration; similarly for down-regulation. Thus, for a fixed false-positive rate of $s\%$, the confidence attained in a positive (resp. negative) level in a pattern depends on the group, i.e. on the position in the pattern, as it depends on Prob(predicted up) for that group.

As a consequence of this approach, when we see a level different from 0, we have a certain confidence in the gene (tag) being up-regulated (if the level is positive) or down-regulated (if the level is negative) as compared to the reference group. However, when we see a 0 there is no confidence implied for the statement that the gene (tag) is not differentially regulated. We can only take 0 to mean that we do not have enough evidence to support a change in level.

Implementation

Software implementing the pattern generation protocol described in the previous section is available from the authors. A brief description of the program follows. Details on usage and input formats are described in the documentation, available from the authors (this documentation also includes a concise summary of our approach).

The program (written in Perl) takes as input a file containing a list of filenames, where each file in the list

gives the results of one experiment. In the input file, the user specifies which experiments are replicates for the same sample type and whether ratios should be taken to the median or to a reference group. Various parameters to be used can be optionally specified, otherwise default values are used. These parameters include the desired false-positive rate $s\%$, a default cutratio to be used when no replicates are available, and either a list of values (called *minimum useful values*), one per sample, or a parameter d to be used in the preprocessing step for certain experiment platforms (e.g. for filter arrays). If the list of minimum useful values is given, this means that the user deems that for each sample only those intensities above the minimum value (s)he provided for that sample are likely to have given hybridization signals above background. The user might specify a single value d instead, if from the knowledge about the experiment procedure, it is deemed that there is a d such that, for any sample in the collection under scrutiny, intensity values which are in the top $d\%$ are those which can be distinguished from background. In the latter case, for each experiment a minimum useful value is then computed using d. Finally a minimum useful value is attached to each group by averaging the minimum useful values of the replicates in that group. In the version of the algorithm given in Figure 2 (version I), only those gene tags with values above the minimum useful value in every group are assigned a pattern. There is another version (version II) of the software which considers every gene tag such that in at least one of the groups its value is above the minimum useful value (and for the groups at which the value is below the minimum useful value, the gene tag's value is raised to be equal to the latter). Thus in version II we account for any signal that might be considered as real while avoiding undue influence of background noise.

The main output of this program consists of an `html` file. The file displays information such as the list of level cutoffs for each group, various counts of interest, and the list of generated patterns. For each such pattern, the gene tags to which that pattern has been assigned are listed. Moreover, the gene tag identifiers in each cluster can be linked to appropriate databases, when specified, to give information on the gene in question (see Section **Results** for an example of this). Sample sections of an output `html` file are shown in Figure 3.

Results

Application to an erythroid development nylon filter dataset

We have used the software described above to analyze filter array experiments of different erythroid development cell samples. These experiments were performed in part to gain insight into the molecular basis for erythroleukemia.

Our erythroid development dataset contains five homotypic groups representing an erythroleukemic cell line and normal cells under different conditions. There are replicate data for each of the groups.

The groups are: CD34 positive cells (human blood progenitor cells including those for erythroid cells; we will write CD34 as a shorthand), human adult and cord erythroblasts (red blood cell precursors), HEL (human erythroleukemia) cells, and HEL cells treated with hemin (induced to express erythroid genes). The HEL group roughly represents a leukemic equivalent of the CD34 group. Similarly, the HEL+hemin group roughly represents a leukemic equivalent of the erythroblast groups (it represents a leukemic cell forced to differentiate and stop growing). The details on the preparation of the cells and on the generation of the data will be published elsewhere along with further biological interpretation of the analysis results (McKenzie *et al.*, manuscript in preparation). Briefly, mRNA was isolated from the cells and reverse-transcribed, then radioactively-labelled by random priming, and interrogated by hybridization to arrays of IMAGE clones using either GenomeSystems GDA filter v1.2 or GDA v1.3 filter 1. The hybridization signals were detected using a Molecular Dynamics Storm PhosphorImager and quantitated by GenomeSystems, or using the Genomic Solutions BioImage software package. Intensities of duplicate spots (on the array) for the same clone (gene tag) were merged into one datum by averaging. Spots which failed visual inspection for artifacts and duplicate spots whose intensities differed by more than two-fold were rejected from further analysis. The signal (clone) intensities were normalized by calculating each signal as percent of the total array signal. Since the sample types in this dataset were closely related cells and a very large number of mRNAs were assessed, it was reasonable to make the assumption that the total mRNA abundance for clones on the filter used would not change considerably from sample to sample. The identifier for each signal was the IMAGE clone ID allowing simple comparisons between the two types of filters used. The IMAGE clone ID was also used to generate names and links for further information on the genes represented on the filter array. More precisely, information on the identity of the gene represented by the clone was provided by DOTS, a database of transcribed sequences (http://www.cbil.upenn.edu/DOTS). The transcribed sequences in DOTS are consensus sequences derived from assemblies of ESTs (including all those available from IMAGE libraries). The transcribed sequences were used to search nucleic acid and protein databases for homology to known genes. The information provided (see Figure 3(b) for an example) is the name of the homologous gene or protein, the database searched, and the percent identity and length of the best high

(a)

The levels for group 1 are: 0, 1, 2
The levels for group 2 are: 0, 1, 2
The levels for group 3 are: 0, 1
The levels for group 4 are: 0, 1, 2

There are 445 gene tags

for group 1 there are 17 up symbols and 0 down symbols
for group 2 there are 7 up symbols and 0 down symbols
for group 3 there are 3 up symbols and 0 down symbols
for group 4 there are 9 up symbols and 0 down symbols

the false positive rate is 0.001

the confidence for the up symbols for group 1 is 0.973823529411765
there are no down symbols for group 1

the confidence for the up symbols for group 2 is 0.936428571428571
there are no down symbols for group 2

the confidence for the up symbols for group 3 is 0.851666666666667
there are no down symbols for group 3

the confidence for the up symbols for group 4 is 0.950555555555556
there are no down symbols for group 4

(b)

Pattern 0,2,1,0 has been attached to the following 3 gene tags:	
245543	135862 1ABY|A Brookhaven Protein Data Bank (nrdb) 66 50 Chain A, Cyanomet Rhb1.1 (Recombinant Hemoglobin) 135862 HT1701 EGAD 95 207 recombination protein A 341567 HT2857 EGAD 88 137 globin, alpha 1
296258	181602 3170176 GenPept 100 114 antigen NY-CO-3
230534	

Fig. 3. Sample sections of the `html` file output by the pattern generating software (version I) applied to the groups: CD34, adult erythroblast, cord erythroblast, and HEL+hemin with ratios to the reference group HEL (described in section 4.1) and parameters $d = 15$ and $s\% = 0.1\%$. (a) Part of the report on the data, containing the list of levels for each group, various counts, and the confidence for each group. (b) Part of the report on one of the generated patterns, displaying the gene tags with that pattern and their descriptions from DOTS, when available.

scoring pair (HSP) using BLAST. No information is provided if no significant matches were found. A link is provided through the IMAGE clone identifier for further information on the DOTS transcribed sequence containing that clone, such as chromosomal map locations for ESTs, cellular roles, source libraries, and protein motifs.

We have applied our pattern-generating software to analyze the groups described above. The available replicates were: two CD34, three adult erythroblasts, two cord blood erythroblasts, three HEL, and two HEL+hemin experiments. With notation as in Section **Implementation**, the value of d was set at 15 reflecting the consideration that only the moderate to highly abundant mRNA classes (greater than or equal to 10 copies per cell, see Zhang *et al.* (1997)) were likely to have given hybridization signals above background on the filter array.

Ratios were generated using the HEL group as reference and the software was run both by merging the adult and cord erythroblasts into one group, the erythroblasts, with five replicates and by keeping them in two separate groups. We merged them when looking for differences between normal and leukemic cells, because the developmental stage was not considered relevant in this case. In the first case, the pattern length is therefore $\ell = 3$ and the three (non-reference) groups were listed in the pattern in the order: CD34, erythroblasts, HEL+hemin. In the second case the pattern length is $\ell = 4$ and the groups were

listed in the order: CD34, adult, cord, HEL+hemin. Both version I and version II of the algorithm were used (see Section **Implementation**). For these data, the running time was always below 90 seconds, when the software was run on a Sun Ultra-10 running an UltraSPARC IIi CPU at 300MHZ with 128MB RAM.

When the adult and cord were merged, out of the 18,123 clones which were present in at least one experiment for each group, 540 were above the minimum useful value in every group and 5063 were above the minimum useful value in at least one group. The value for s was varied. The higher the s value, the richer was the expression description and the lower the confidence in our predictions, when version I was used. (As noted in Section **Methods and Algorithm** however, the confidence depends also on the probability of predicting a gene tag as up-regulated in a group. Therefore, in general, lowering the false positive rate does not necessarily cause an increase in confidence.) For $s\% = 1\%$ there are 5 levels for CD34 (from 0 to 4), 10 levels for the erythroblasts (from -1 to 8), and 6 levels for HEL+hemin (from -1 to 4). When $s\%$ is lowered to 0.1% there are three levels (from 0 to 2) for each of these three groups. When $s\%$ is lowered again to 0.01% there are 2 levels (0 and 1) for each of these groups. The following table illustrates how the confidence in predicted up-regulation changes with s and with the group. After each group the number (out of 540) of clones up-regulated in that group as compared to the reference is also displayed (denoted by '#')

$s\%$	CD34 conf.	#	ery. conf.	#	HEL+h. conf.	#
1%	85%	37	81%	28	89%	47
0.1%	97%	17	92%	7	94%	9
0.01%	98%	3	99%	5	99%	4

Clones representing the same gene were usually found to have identical or very similar patterns, as expected. Furthermore, clones representing genes whose expression is known in these cells presented patterns compatible with what was expected. For example, when version I of the software was run keeping adult and cord separate, the α- and the β-globin clones whose raw intensities are represented in Figure 4, were assigned the patterns $(0, 8, 2, 0)$ and $(0, 8, 1, 0)$ respectively, with moderate to high confidence. Both clones were therefore detected as up-regulated in erythroblasts and more in adult than in cord. This matches what is known about the expression of these genes (see Papayannopoulou *et al.* (1987), Dalyot *et al.* (1992), and Ni *et al.* (1999)), for example the fact that β globin is replaced with fetal γ globin in cord samples.

We can ask what genes are differentially expressed between normal (CD34, erythroblasts) and leukemic cells. From that set, we can then ask which genes are induced by hemin (HEL+hemin) to adopt a normal expression pattern, to be followed up by further experiments. The experiments were analyzed for this purpose using both version I and version II of the software. Different false-positive rates were applied to each case to achieve moderate to high confidence in the assertions of differential expression. These rates for high confidence, and the genes identified as differentially expressed for both cases, are presented in the table in Figure 5. Having more genes available to start with (5063 vs. 540) led to more genes identified as differentially expressed but at lower confidence. At similar confidence levels, starting with more genes did not necessarily lead to more genes identified as differentially expressed between normal and HEL cells as can be seen from the table. There was general agreement for both cases and several candidate genes were identified for further investigation. These include, at moderate confidence, a member of a signal transduction cascade (MAPKK2). Functional studies are required for the next step of this analysis.

Comparison with standard statistical analysis methods

Since the distributions of gene intensities for the groups in the dataset of Section **Application to an erythroid development nylon filter dataset** can be reasonably assumed to be close to normal, we have applied standard statistical analysis methods to this dataset, to compare the results with those of our protocol (we used version I). The standard methods consist in combining tests like the t-test with a 'Bonferroni-like' correction. The Bonferroni correction in itself is too strict, because it is generally used to ensure at high confidence the absence of false positives (see Claverie, 1999). However, a similar correction can be applied to ensure at high confidence that the number of false positives is no larger than $s\% \times$ (number of gene tags). (For example, this correction can be done using a Poisson distribution with parameter (number of gene tags)$\times p$, whereby the significance threshold p that one needs to use in a t-like test done on a gene tag by gene tag basis can be determined.) This puts the standard methods on a footing which allows a fair comparison with our method (further details about this can be found in the documentation available from the authors).

The shift used in our method to be able to apply approximation (4) has certain consequences that should be pointed out. Namely, it causes genes with lower expression levels to require stronger evidence for differential expression than those with higher levels. This is reasonable since, for example, a change from 1 mRNA

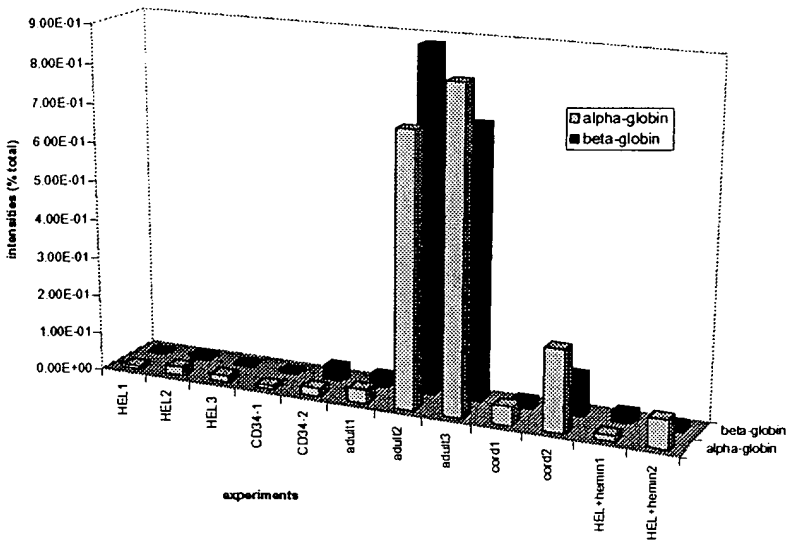


Fig. 4. Input intensity values for an α- and a β-globin clone in the erythroid development dataset. The lists of ratios of the average unshifted intensities of the non-reference groups (CD34, adult, cord, HEL+hemin) to the average unshifted intensity of the HEL (reference) group for the α- and β-globin clones were respectively: (0.965, 34.1, 8.49, 2.93) and (2.51, 66.6, 7.74, 2.5).

copy per cell in group 0 to 2 in group i is less compelling evidence for up-regulation in group i than a change from 100 to 200.

Both for the dataset where adult and cord erythroblasts were merged and for the one where they were kept separate, we compared our results for the various values of $s\%$ used with results obtained by running the standard methods. The comparison was done group by group (in other words, position by position in the patterns). For $s\% = 0.1\%$ and $s\% = 0.01\%$ (such low values are not unreasonable if one desires good confidence in the predictions) the standard methods did not detect any differentially expressed gene tags (as compared to the reference group HEL) in any position. Not even the α- and the β-globin were detected as up-regulated in erythroid cells, as they are known to be. In fact, even for $s\% = 1\%$ the latter were not detected as up-regulated by the standard methods. For $s\% = 1\%$ no gene tags were detected by the standard methods as differentially expressed in any of CD34, or the erythroblasts, whereas our method detected several at high confidence. For the HEL+hemin group the standard methods detected about 20 gene tags as upregulated with high confidence and we detected about 50 with even higher confidence. There was no overlap between these two sets. The gene tags detected by the standard methods had low intensities (as compared to our numerical shift) and we expected to miss those for the reasons explained above. (In Section **Discussion** we pick this point up again in reference to future work.)

Application to a two-channel microarray lymphocyte dataset

We have applied our protocol to analyze some of the experiments published by Alizadeh *et al.* (2000). These consist of two-channel microarray experiments, which use a specialized microarray, named 'Lymphochip.' The following nine experiments were downloaded from http://llmpp.nih.gov/lymphoma/data/rawdata/ and used in our analysis: three samples of normal blood B-cells (lc7b023, lc7b024, lc7b103), four samples of CLL (B-cell chronic lymphocytic leukemias: lc7b047, lc7b048, lc7b069, lc7b070), and two samples of FL (B-cells from follicular lymphoma: lc7b096, lc7b097). This subset was chosen because it provided replicates using the same array version (lc7b) for three closely related groups. For each experiment, only those spots on the array which were not 'flagged' as bad and which were above background (total signal $\geq 1.4 \times$ background, fraction of pixels above background ≥ 0.55 for both channels) were retained, in accordance with the authors own analysis (the resulting number of gene tags was 5595.) The median ratios (MRAT) output from the ScanAlyze software program (http://rana.stanford.edu/software/) relative to a common control sample (a pool of nine lymphoma cell lines) were used to attach a value to each clone (gene tag) by averaging over the spots representing that clone. Finally, for each experiment, the list of values attached to the clones were rescaled to have a median of 1 (this is similar to what was done by Alizadeh *et al.* (2000)). The values

IMAGE clone ID	description	version I $s\% = 0.25\%$	version II $s\% = 0.01\%$
198960	unknown, down in CD31		√
296266	unknown, down in CD34		√
241804	oncogene tre-2, down in CD34		√
144221	β-globin, up in erythroblasts		√
118125		√	√
136255		√	
211976	α-globin, up in erythroblasts	√	√
215513		√	√
296258	NY-CO-3 antigen, up in erythroblasts	√	√
306759	similar to α-2, 3-sialyltransferase, up in erythroblasts		√
230534	unknown, up in erythroblasts	√	√
201839	unknown, up in erythroblasts		√
189110	unknown, up in erythroblasts	√	
215162	ferritin, up in erythroblasts and HEL+hemin	√	
191153	unknown, up in CD31	√	√
116131	unknown, up in CD31	√	√
267179	ribosomal protein S3, up in CD31	√	
203939	splicing factor, up in CD31	√	
233938	kinetochore motor CENP-E, up in CD31	√	
233993	similar to adipogenesis-inhibitory factor, up in CD31	√	
316621	similar to myosin, heavy polypeptide embryonic, up in CD31	√	
116127	similar to RNA helicase, up in CD31	√	
197281	unknown, up in CD31	√	
36332	unknown, up in CD31	√	
36118	unknown, up in CD34	√	
179819	unknown, up in CD31	√	
191910	unknown, up in CD34	√	
193802	unknown, up in CD34	√	
491451	ribosomal protein L27a, up in CD34	√	
207962	similar to H326, up in CD34	√	
210513	KIAA0381, up in CD31	√	
258673	unknown, up in CD31	√	
36821	unknown, up in CD31	√	
317522	unknown, up in CD31	√	
311125	histone H2A.2, up in CD31 and HEL+hemin	√	
171861	similar to neuroD, up in CD31 and HEL+hemin	√	
222633	unknown, up in CD31 and HEL+hemin	√	√

Fig. 5. Differentially-expressed genes between normal (CD34, erythroblasts) and HEL cells. Genes altered upon hemin induction (HEL+hemin) to a normal pattern are indicated. The checkmarks denote those clones which had the differential expression specified in the description column. With version I of the software and a false-positive rate of 0.25%, the confidence for up-regulation in the CD34 group was 94% (no down-regulation), the confidence for up-regulation in the erythroblast group is 83% (no down-regulation), and the confidence for up-regulation of the HEL+hemin group is 78% (no down-regulated genes with acceptable confidence). With version II of the software and a false-positive rate of 0.01%, the confidence for the CD34 group was 83% for up-regulation and 83% for down-regulation, the confidence for the erythroblast group was 93% for up-regulation (no down-regulation), and the confidence for the HEL+hemin group was 87% for up-regulation (no down-regulation).

obtained in this way were the normalized intensities input into our program.

Our program found at high confidence differentially expressed genes between the lymphoma (CLL, FL) and normal blood B-cells despite their high overall similarity in expression profiles (Alizadeh *et al.*, 2000). It also found germinal B-cell associated genes (BCL-6, A-myb) to be up-regulated only in FL and it did not detect cell proliferation genes as differentially expressed in either CLL or FL. These results are in accordance with the major findings for these sample types in Alizadeh *et al.* (2000). In addition, we found (but they did not report) that fos and jun transcription factors (which form the AP-1 complex) were down-regulated at high confidence (93%)

in CLL alone. The inability of tumor necrosis factor (TNF) to induce c-fos and c-jun in CLL cells has been linked to the refractory nature of CLL to stimulation of cell proliferation (Jabbar *et al.*, 1994).

Our approach therefore provides results which are consistent overall with results obtained from clustering and visual inspection methods and at the same time allows us to attach confidence to the predictions made about up- or down-regulation. Moreover it provides new insights into differentially expressed genes.

Finally, we note that standard statistical analysis methods cannot in general be applied to situations when the comparisons are between red-to-green ratios from two separate microarrays (like the situation of these data), as the assumption that the intensities so normalized have a close to normal distribution is no longer valid.

REMARK. For both the dataset of Section **Application to an erythroid development nylon filter dataset** and that of Section **Application to a two-channel microarray lymphocyte dataset** we deemed that the normalized intensities fit in the framework of applicability of approximation (4). For the former dataset, they could be reasonably assumed to have distributions close to normal. For the second, they could be assumed to have distributions which were not highly asymmetric, because of the way the gene tags were selected (denominators in the MRATs having distributions concentrated away from 0).

Discussion

We wanted to formally address the issue of generating descriptive, yet reliable gene expression patterns for datasets containing a few replicate (hybridization array) experiments per sample type and this paper presents a novel protocol in this direction.

Some investigators choose to take the position that there is so much noise in this kind of data that the best that can be done is to divide genes (for each group) into two categories (on or off); i.e. they choose a binary representation of gene expression. This approach seems overly pessimistic and will likely cause one to miss real and interesting changes in expression levels that are represented reliably in the data. Also this approach does not allow for comparisons between different sample types if a gene is turned on in both types. When possible, we give a representation of expression levels which is richer than a binary one. The levels are discrete (as we bin according to the variability of the data). Our protocol will in fact assign binary representations if the data in question are sufficiently noisy to merit that. In the most extreme case, our protocol might even assign only a *single* level to one or more sample types, in case the data are so variable that they cannot give any reliable indications of differential expression.

It is desired to generate patterns in such a way as to have a certain degree of confidence in the predictions made for up- and down-regulation. When many replicates are available, standard (or quasi-standard) statistical methods can be used to detect differentially expressed genes for certain kinds of data (see for example Claverie (1999) and Golub *et al.* (1999)). These methods involve measures of variability that depend on the sample type as well as on the gene. However, for various reasons, chief among which is the current high cost of the experiments, many laboratories generate data which do not have many replicates. Thus one needs to get as much as possible out of just a few replicates. Moreover the ultimate goal is not just to have a false-positive rate on the order of 5% or even 1%, but to have a high confidence in the predictions made. In other words we desire a low probability that a gene is not up- (resp. down-) regulated, given that we predict it is. Since the percentage of differentially expressed genes in two sample types is usually low (in closely related cells this is expected to be 1.5% to 2.5%, see (Sagerstrom *et al.*, 1997)), false-positive rates much lower than these are needed to obtain reasonable confidences. Our method was developed with these two aspects in mind. While few replicates are needed, it should be stressed that, without any replicates (as in the case of a number of recently published two-color microarray studies), one cannot determine the false-positive rate as a function of the criteria by which differential expression is predicted, making a judicious choice of cutoffs difficult to impossible.

It is important to note that the issue at hand is one of gaining confidence in differential expression. Therefore, in the patterns generated with our protocol, when one sees an expression level at a group of 1 or more (resp. -1 or less), there is a certain degree of confidence that the gene is up-regulated (resp. down-regulated) in this group as compared to the reference group. However, if there is an expression level of 0, there is not the same degree of confidence that it is *not* up- or down-regulated as compared to the reference. It is not possible to have confidence about both phenomena at the same time without losing some information. One could, however, develop methods to extract confidence measures of non-differential expression. We have not done so here, but will investigate this in future work.

In the applications discussed in Sections **Application to an erythroid development nylon filter dataset** and **Comparison with standard statistical analysis methods**, our method picked up differentially expressed genes which were not picked up by the standard methods at comparable confidence and which were known to be differentially expressed. In Section **Comparison with**

standard statistical analysis methods we noted that, as an effect of the numerical shift used in our approach to apply approximation (4), genes whose expression is very low as compared to the shift require stronger evidence before we detect differential expression. In some of the runs performed there were a few genes that the standard methods declared as differentially expressed and that we did not. This only occurred though when $s\% = 1\%$ and never occurred for lower values of s. These tended to be genes in the low intensity category (as compared to our shift), which we were not expecting to detect. Plans for future work include the development of protocols to apply our method iteratively to different intensity level subsets of the gene tags under consideration to pick up those genes in the low intensity category that might be true positives and that are missed on the first run (we have some promising preliminary results in this respect). For the cases where the standard methods are applicable, we are also considering ways of combining our method with the standard methods so to lower the false-negative rate of our predictions.

As a final remark, we note that the input to our program consists of normalized intensities and it is up to the user to choose an appropriate normalization based on the way the data have been generated. Moreover, it is also important that the intensity distributions (after normalization) are appropriate for approximation (4) (see Section **Methods and Algorithm** and Remark in Section **Results**).

In summary, we have developed a novel protocol to assign measures of confidence to differentially expressed genes. The protocol takes advantage of the large number of genes typical of hybridization array experiments to control for the variability in gene expression values, using few replicates in an approach that provides low false-positive rates and high confidence. The protocol goes beyond a simple binary description of expression, when justified, and provides discrete descriptions of gene expression patterns based on the desired level of confidence.

Note: Software documentation and further discussion of issues relative to this paper can be found at http://www.cbil.upenn.edu/PaGE.

Appendix

To justify the approximation of the distribution of $\frac{\overline{X}_{g,j}}{\mu_{g,j}}$ $(j = 0, i)$, for g varying over the genes, by the distribution of (4), in the case in which the gene intensities are normally distributed and the distribution of $\rho_{g,j} = (\sqrt{t_j})\mu_{g,j}/\sigma_{g,j}$ (for g varying over the genes) is concentrated over sufficiently large values, we use the following lemma.

LEMMA 1. *Let $X_1, X_2, \ldots, X_t$ be a random sample of a normal random variable X with positive mean μ_X and standard deviation σ_X. Let $\overline{X} = (X_1 + X_2 + \cdots + X_t)/t$. Let $k \in \{1, 2, \ldots, t\}$. Then,*

$$\frac{\frac{X_k}{\overline{X}} - 1}{\sqrt{t-1}} = \frac{Z_1}{\rho_X + Z_2} \tag{6}$$

and

$$\frac{\overline{X}}{\mu_X} - 1 = \frac{Z_2}{\rho_X}, \tag{7}$$

where Z_1 and Z_2 are two independent standard normal random variables and $\rho_X = (\sqrt{t})\mu_X/\sigma_X$.

PROOF. From the properties of sums of random variables it is easy to check that the random variables $X_k - \overline{X}$ and $\overline{X}$ are independent and normally distributed with respective means 0 and μ_X and respective standard deviations $\left(\sqrt{\frac{t-1}{t}}\right)\sigma_X$ and $\frac{1}{\sqrt{t}}\sigma_X$ (to check independence it is sufficient to check that the correlation is 0 since the two random variables are normal). So

$$X_k - \overline{X} = \left(\sqrt{\frac{t-1}{t}}\right)\sigma_X Z_1 \tag{8}$$

and

$$\overline{X} = \frac{1}{\sqrt{t}}\sigma_X Z_2 + \mu_X = \frac{\sigma_X Z_2 + (\sqrt{t})\mu_X}{\sqrt{t}}. \tag{9}$$

Dividing both sides of (8) by $\sqrt{t-1}$, then taking the ratio of (8) and (9) and dividing top and bottom of the resulting right hand side by σ_X, we get (6). Since $\overline{X}$ is normal with positive mean μ_X and standard deviation $\frac{1}{\sqrt{t}}\sigma_X$, $\frac{\overline{X}}{\mu_X} - 1$ is normal with mean 0 and standard deviation $1/\rho_X$, thus (7) follows. □

With notation as in Section **Methods and Algorithm**, if we fix a homotypic group, say group j and we apply this lemma to $X = X_{g,j}$, we get that

$$\frac{\frac{X_{g,j,k}}{\overline{X}_{g,j}} - 1}{\sqrt{t_j - 1}} = \frac{Z_1}{\rho_{g,j} + Z_2} \tag{10}$$

and

$$\frac{\overline{X}_{g,j}}{\mu_{g,j}} - 1 = \frac{Z_2}{\rho_{g,j}}. \tag{11}$$

Therefore, if the distribution of $\rho_{g,j}$ (for g varying over the genes) is concentrated away from zero, the distribution of (10) will be a very good approximation to that of (11). When $\rho_{g,j}$ is around 8 the approximation is extremely close.

Acknowledgments

We thank Warren Ewens for many useful conversations and Eric Slud for suggesting the proof given in the **Appendix**. We also thank Brian Brunk for the use of DOTS; Hong Ni, Christopher Orr and Linda Schmidt for their collaboration in generating the erythroid development biological data; and Ash Alizadeh for providing the mapping of clones to array elements for their data. Finally, we thank the referees for their comments.

This work has been supported in part by the NSF training grant BIR 9413215, the NIH grants RO1-RR-04026 and N01 CN 95037, and by the Nemours foundation.

References

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Claverie,J.M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet. Sci.*, **8**, 1821–1832.

Chen,Y., Dougherty,E.R. and Bittner,M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics*, **2**, 364–374.

Dalyot,N., Fibach,E., Rachmilewitz,E.A. and Oppenheim,A. (1992) Adult and neonatal patterns of human globin gene expression are recapitulated in liquid cultures. *Exp. Hematol.*, **20**, 1141–1145.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Jabbar,S.A., Hoffbrand,A.V. and Gitendra Wickremasinghe,R. (1994) Regulation of transcription factors NF kappa B and AP-1 following tumour necrosis factor-alpha treatment of cells from chronic leukaemia patients. *Br. J. Haematol.*, **86**, 496–504.

Ni,H., Yang,X.D. and Stoeckert,C.J.,Jr (1999) Maturation and developmental stage-related changes in fetal globin gene expression are reproduced in transiently transfected primary adult human erythroblasts. *Exp. Hematol.*, **27**, 46–53.

Papayannopoulou,T., Nakamoto,B., Kurachi,S. and Nelson,R. (1987) Analysis of the erythroid phenotype of HEL cells: clonal variation and the effect of inducers. *Blood*, **70**, 1764–1772.

Sagerstrom,C.G., Sun,B.I. and Sive,H.L. (1997) Subtractive cloning: past, present, and future. *Ann. Rev. Biochem.*, **66**, 751–783.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Glolub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, **96**, 2907–2912.

Wen,X., Fuhrman,S., Michaels,G.S., Carr,D.B., Smith,S., Barker,J.L. and Somogyi,R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA*, **95**, 334–339.

Zhang,L., Zhou,W., Velculescu,V.E., Kern,S.E., Hruban,R.H., Hamilton,S.R., Volgestein,B. and Kinzler,K.W. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.

Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

Ash A. Alizadeh[1,2], Michael B. Eisen[2,3,4], R. Eric Davis[5], Chi Ma[6], Izidore S. Lossos[6], Andreas Rosenwald[5], Jennifer C. Boldrick[1], Hajeer Sabet[5], Truc Tran[5], Xin Yu[5], John I. Powell[7], Liming Yang[7], Gerald E. Marti[8], Troy Moore[9], James Hudson Jr[9], Lisheng Lu[10], David B. Lewis[10], Robert Tibshirani[11], Gavin Sherlock[4], Wing C. Chan[12], Timothy C. Greiner[12], Dennis D. Weisenburger[12], James O. Armitage[13], Roger Warnke[14], Ronald Levy[6], Wyndham Wilson[15], Michael R. Grever[16], John C. Byrd[17], David Botstein[4], Patrick O. Brown[1,18] & Louis M. Staudt[5]

Departments of [1]Biochemistry, [3]Genetics, [14]Pathology, [6]Medicine, [10]Pediatrics and [11]Health Research & Policy and Statistics, and [18]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA

[5] Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

[7] Bioinformatics and Molecular Analysis Section, CBEL, CIT, NIH, Bethesda, Maryland 20892, USA

[8] CBER, FDA, Bethesda, Maryland 20892, USA

[9] Research Genetics, Huntsville, Alabama 35801, USA

Departments of [12]Pathology and Microbiology, and [13]Internal Medicine, University of Nebraska Medical Center, Omaha, Nebraska 68198, USA

[15] Medicine Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

[16] Johns Hopkins Oncology Center, Johns Hopkins School of Medicine, Baltimore, Maryland 21287, USA

[17] Walter Reed Army Medical Center, Washington, DC 20307, USA

[2] These authors contributed equally to this work

Diffuse large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgkin's lymphoma, is clinically heterogeneous: 40% of patients respond well to current therapy and have prolonged survival, whereas the remainder succumb to the disease. We proposed that this variability in natural history reflects unrecognized molecular heterogeneity in the tumours. Using DNA microarrays, we have conducted a systematic characterization of gene expression in B-cell malignancies. Here we show that there is diversity in gene expression among the tumours of DLBCL patients, apparently reflecting the variation in tumour proliferation rate, host response and differentiation state of the tumour. We identified two molecularly distinct forms of DLBCL which had gene expression patterns indicative of different stages of B-cell differentiation. One type expressed genes characteristic of germinal centre B cells ('germinal centre B-like DLBCL'); the second type expressed genes normally induced during *in vitro* activation of peripheral blood B cells ('activated B-like DLBCL'). Patients with germinal centre B-like DLBCL had a significantly better overall survival than those with activated B-like DLBCL. The molecular classification of tumours on the basis of gene expression can thus identify previously undetected and clinically significant subtypes of cancer.

Despite the variety of clinical, morphological and molecular parameters used to classify human malignancies today, patients receiving the same diagnosis can have markedly different clinical courses and treatment responses. The history of cancer diagnosis has been punctuated by reassortments and subdivisions of diagnostic categories. There is little doubt that our current taxonomy of cancer still lumps together molecularly distinct diseases with distinct clinical phenotypes. Molecular heterogeneity within individual cancer diagnostic categories is already evident in the variable presence of chromosomal translocations, deletions of tumour suppressor genes and numerical chromosomal abnormalities. The classification of human cancer is likely to become increasingly more informative and clinically useful as more detailed molecular analyses of the tumours are conducted.

The classification of human lymphomas has steadily evolved since their initial recognition by Thomas Hodgkin in 1832 (ref. 1). Beginning with the distinction of Hodgkin's disease from other malignant and non-malignant conditions[2,3], a variety of lymphoma classifications have been advanced on the basis of both morphologic and molecular parameters[4]. The most recent classification scheme, the Revised European–American Lymphoma (REAL) classification, was introduced to categorize distinct clinical–pathological entities[5]. However, within this classification system, various morphologic subtypes were unified into groups despite the suspicion that they "include more than one disease entity"[5].

Diffuse large B-cell lymphoma (DLBCL) is one disease in which attempts to define subgroups on the basis of morphology have largely failed owing to diagnostic discrepancies arising from inter- and intra-observer irreproducibility[5,6]. Diffuse large B-cell lymphoma is an aggressive malignancy of mature B lymphocytes, with an annual incidence of over 25,000 cases, accounting for roughly 40% of cases of non-Hodgkin's lymphoma. Patients with DLBCL have highly variable clinical courses: although most patients respond initially to chemotherapy, fewer than half of the patients achieve a durable remission[6,7]. Although a combination of clinical parameters is currently used to assess a patient's risk profile, these prognostic variables are considered to be proxies for the underlying cellular and molecular variation within DLBCL[8].

An important component of the biology of a malignant cell is inherited from its non-transformed cellular progenitor. Each of the currently recognized categories of B-cell malignancy has been tentatively traced to a particular stage of B-cell differentiation, although the extent to which these malignancies maintain the molecular and physiological properties of normal B-cell subsets is not clear. The rearranged immunoglobulin genes in DLBCL and most other non-Hodgkin's lymphomas bear mutations that are characteristic of somatic hypermutation, an antibody-diversification

[4] Present address: Life Sciences Division, Lawrence Orlando Berkeley National Labs and Department of Molecular and Cellular Biology, University of California, Berkeley, California 94720, USA.

mechanism that normally occurs only within the germinal centre of secondary lymphoid organs[9]. This evidence suggests that DLBCL arises either from germinal centre B cells or from B cells at a later stage of differentiation.

Here we examined the extent to which genomic-scale gene expression profiling can further our understanding of B-cell malignancies. We addressed whether we could (1) generate a molecular portrait of distinct types of B cell malignancy; (2) identify distinct types of B-cell malignancy not recognized by the current classification system; and (3) relate each malignancy to normal stages in B-cell development and physiology. We focused particularly on DLBCL to determine whether gene expression profiling could subdivide this clinically heterogeneous diagnostic category into molecularly distinct diseases with more homogeneous clinical behaviours.

Construction of a specialized DNA microarray

Recent technical and analytical advances make it practical to quantitate the expression of thousands of genes in parallel using complementary DNA microarrays[10]. This mode of analysis has been used to observe gene expression variation in a variety of human tumours[11-17]. To apply this method to questions in normal and malignant lymphocyte biology, we designed a specialized microarray—the 'Lymphochip'—by selecting genes that are preferentially expressed in lymphoid cells and genes with known or suspected roles in processes important in immunology or cancer[18].

Because of the suspected importance of the germinal centre B cell to the genesis of non-Hodgkin's lymphomas, 12,069 out of the 17,856 cDNA clones on this microarray were chosen from a germinal centre B-cell library[18]. An effort was made to include all distinct genes that were initially discovered in this library. We included an additional 2,338 cDNA clones from libraries derived from DLBCL, follicular lymphoma (FL), mantle cell lymphoma and chronic lymphocytic leukaemia (CLL). Finally, we added clones representing a variety of genes that are induced or repressed during B- and T- lymphocyte activation by mitogens or cytokines[19] and a curated set of 3,186 genes of importance to lymphocyte and/or cancer biology. About a quarter of the genes included in this microarray were represented by two or more different cDNA clones, providing internal controls for the reproducibility of gene expression quantitation. See Supplementary Information for the complete annotated list of these cDNAs.

Analysis of gene expression in lymphoid malignancies

We used these microarrays to characterize gene expression patterns in the three most prevalent adult lymphoid malignancies: DLBCL, FL and CLL (Fig. 1). To provide a framework for interpretation of the gene expression in these patient samples, we also profiled gene expression in purified normal lymphocyte subpopulations under a range of activation conditions, in normal human tonsil and lymph node, and in a variety of lymphoma and leukaemia cell lines. Fluorescent cDNA probes, labelled with the Cy5 dye, were prepared from each experimental messenger RNA sample. A reference cDNA probe, labelled with the Cy3 dye, was prepared from a pool of mRNAs isolated from nine different lymphoma cell lines. Each Cy5-labelled experimental cDNA probe was combined with the Cy3-labelled reference probe and the mixture was hybridized to the microarray. The fluorescence ratio was quantified for each gene and reflected the relative abundance of the gene in each experimental mRNA sample compared with the reference mRNA pool. The use of a common reference probe allowed us to treat these fluorescent ratios as measurements of the relative expression level of each gene across all of our experimental samples.

In all, ~1.8-million measurements of gene expression were made in 96 normal and malignant lymphocyte samples using 128 Lymphochip microarrays. Figure 1 provides an overview of the variation in gene expression across these samples. A hierarchical clustering algorithm was used to group genes on the basis of similarity in the pattern with which their expression varied over all samples[20]. The same clustering method was used to group tumour and cell samples on the basis of similarities in their expression of these genes. The data are shown in a matrix format, with each row representing all the hybridization results for a single cDNA element of the array, and each column representing the measured expression levels for all genes in a single sample. To visualize the results, the expression level of each gene (relative to its median expression level across all samples) was represented by a colour, with red representing expression greater than the mean, green representing expression less than the mean, and the colour intensity representing the magnitude of the deviation from the mean[20].

Distinct clones representing the same gene were typically clustered in adjacent rows in this gene map, indicating that these genes have characteristic and individually distinct patterns of expression and showing that the effects of experimental noise or artefact are negligible. Likewise, where different tumour samples from the same patient were analysed, they were invariably found clustered in immediately adjacent columns. For example, in three cases of FL in which the malignant cells were separated from the normal host cells by magnetic cell sorting, the purified and unpurified samples from the same patient clustered next to each other. Two samples of leukaemic cells from the same CLL patient were obtained 18 months apart, and these samples were more highly correlated in gene expression with each other than with any other patient's CLL cells. The observed patterns of gene expression thus reflected intrinsic differences between the tumours, rather than variation in handling or experimental artefacts. Moreover, these results show that even within a diagnostic category, each cancer patient has a unique tumour with a characteristic gene expression profile.

Figure 1 paints a complex, but remarkably ordered, picture of the variation in gene expression patterns in lymphoid malignancies, with large sets of genes displaying coordinate expression in related biological samples. Although no information on the identity of the samples was used in the clustering, the algorithm segregated, with few exceptions, the recognized classes of lymphoid malignancies based on global similarities in gene expression patterns. Examination of the coordinately expressed genes in each of the B-cell malignancies and comparison with the normal lymphocyte cell populations yielded considerable insights into the biology of these malignancies. The coloured bars on the right of Fig. 1 indicate clusters of coordinately expressed genes that we operationally defined as gene expression 'signatures'. A gene expression signature was named by either the cell type in which its component genes were expressed (for example, the 'T-cell' signature) or the biological process in which its component genes are known to function (for example, the 'proliferation' signature). Thus, the overall gene expression profile of a complex clinical sample such as a DLBCL lymph-node biopsy can be understood, in a first approximation, as a collection of gene expression signatures that reveal different biological features of the sample.

Gene expression patterns and tumour phenotype

One of clearest distinctions between the gene expression patterns of the three B-cell malignancies involved genes that vary in expression with cellular proliferation rates. Both CLLs and FLs were clustered next to resting B-cell samples, which reflects, in part, the fact that both of these malignancies are relatively indolent, with very low proliferation rates. Correspondingly, the genes that define the proliferation signature were not highly expressed in these malignancies (Fig. 2). This gene expression signature included diverse cell-cycle control genes, cell-cycle checkpoint genes, DNA synthesis and replication genes, and the gene Ki67, commonly used to gauge the 'proliferation index' of a tumour biopsy, as previously noted[15]. In general, the more rapidly proliferating DLBCLs had higher expression of the genes in the proliferation signature. Nonetheless, marked differences in the expression of these genes were evident

between individual DLBCL samples, corresponding to the variability in proliferation index that has been previously observed in DLBCL[21].

The most prominent distinction between CLL and FL came from genes that are characteristic of germinal centre B cells (Fig. 2). An extensive cluster of genes distinguished germinal centre B cells from both resting blood B cells and *in vitro* activated blood B cells. This is remarkable because the stimuli used to activate the blood B cells were chosen to mimic those known to be important for germinal centre formation: crosslinking of the immunoglobulin receptor and CD40 signalling. However, it has thus far not been possible to mimic exactly the germinal centre phenotype *in vitro*, as determined by the failure of a variety of activation conditions to induce the expression of BCL-6 protein, a highly specific marker for germinal centre B cells[22]. The germinal centre B-cell gene expression signature shows that germinal centre B cells represent a distinct stage of B-cell differentiation and not merely one specific form of B-cell activation. Support for this notion comes from the fact that the characteristic gene expression program of germinal centre B cells was maintained in a cultured DLBCL cell line in the absence of the germinal centre microenvironment (Figs 1 and 2).

The observation that FLs show a pattern of ongoing somatic hypermutation of immunoglobulin genes has led to the suggestion that the transformation event leading to FL occurs while the B cell is in the germinal centre microenvironment[23]. The gene expression signature of germinal centre B cells was reproduced virtually unchanged in FL, supporting the view that this lymphoma arises from this stage of B-cell differentiation (Fig. 2).

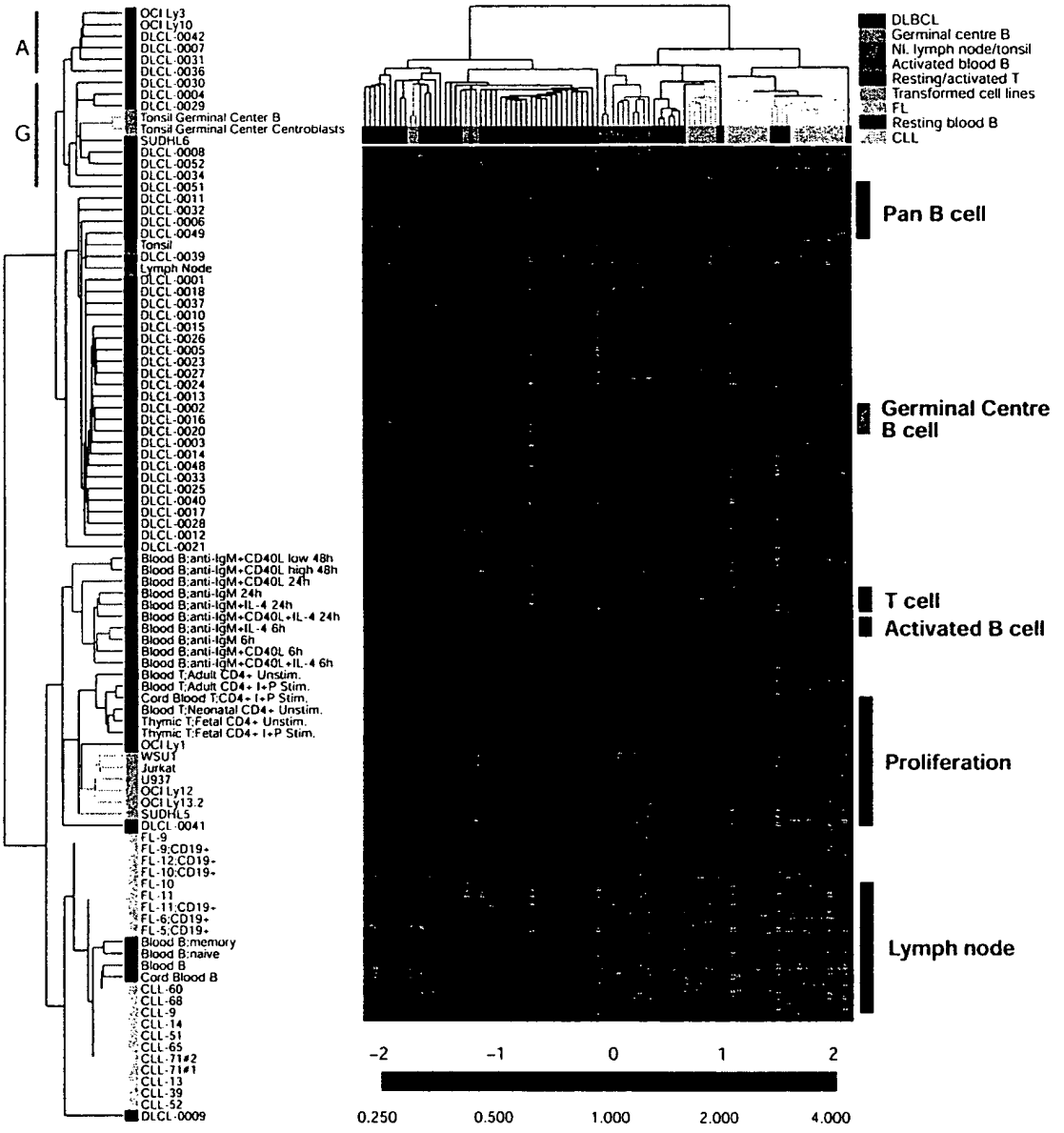


Figure 1 Hierarchical clustering of gene expression data. Depicted are the ~1.8 million measurements of gene expression from 128 microarray analyses of 96 samples of normal and malignant lymphocytes. The dendrogram at the left lists the samples studied and provides a measure of the relatedness of gene expression in each sample. The dendrogram is colour coded according to the category of mRNA sample studied (see upper right key). Each row represents a separate cDNA clone on the microarray and each column a separate mRNA sample. The results presented represent the ratio of hybridization of fluorescent cDNA probes prepared from each experimental mRNA samples to a reference mRNA sample. These ratios are a measure of relative gene expression in each experimental sample and were depicted according to the colour scale shown at the bottom. As indicated, the scale extends from fluorescence ratios of 0.25 to 4 (−2 to +2 in log base 2 units). Grey indicates missing or excluded data. See Supplementary Information for full data.

The gene expression profiles of DLBCLs were largely distinct from those of CLL and FL and showed additional biological complexity in these biopsy samples. Prominent features of the DLBCL profiles appeared to reflect the non-malignant cells in these tumours. A large group of genes defined a 'lymph-node' signature which was shared by most of the DLBCLs and samples of normal lymph node and tonsil (Fig. 2). This signature featured genes encoding known markers of monocytes and macrophages (CD14, CD105, CSF-1 receptor) and natural killer cells (NK4). In addition, genes involved in the remodelling of the extracellular matrix were abundantly expressed (MMP9 matrix metalloproteinase and TIMP-3). All but one DLBCL biopsy displayed the lymph-node signature, but the intensity of this signature varied, possibly reflecting the relative proportion of tumour and host cells in the lymph-node biopsy.

The variable presence of T lymphocytes in DLBCL biopsies was readily discernible by a T-cell gene expression signature that featured components of the T-cell receptor (TCR-β, CD3ε) and genes downstream of T-cell receptor signalling (fyn, LAT, PKC-θ) (Fig. 2). Although this T-cell expression signature was readily apparent in some DLBCLs, it was virtually undetectable in others.

Discovery of DLBCL subtypes

The structure of the hierarchical dendrogram in Fig. 1 indicated that gene expression patterns in DLBCLs might be inhomogeneous. Three branches of the dendrogram captured most of the DLBCLs with only three outlying samples. Clearly, the position of any given DLBCL sample in the dendrogram is determined in a complicated fashion by the influences of several distinct biological themes that are reflected in the expression pattern. Inspection of the gene expression map shown in Fig. 1 suggested that several independent sets of genes were responsible for much of the DLBCL substructure. The expression signatures related to proliferation, T cells and lymph-node biology were differentially represented in the three DLBCL branches. In addition, we noted that the genes that distinguished germinal centre B cells from other stages in B-cell ontogeny were also differentially expressed among DLBCLs, suggesting that B-cell differentiation genes may also be used to subdivide DLBCL. The expression of the germinal centre B cell genes among DLBCLs varied independently from the expression of genes in the other gene expression signatures (Fig. 2; see Supplementary Information for details). In principle, each of these gene expression

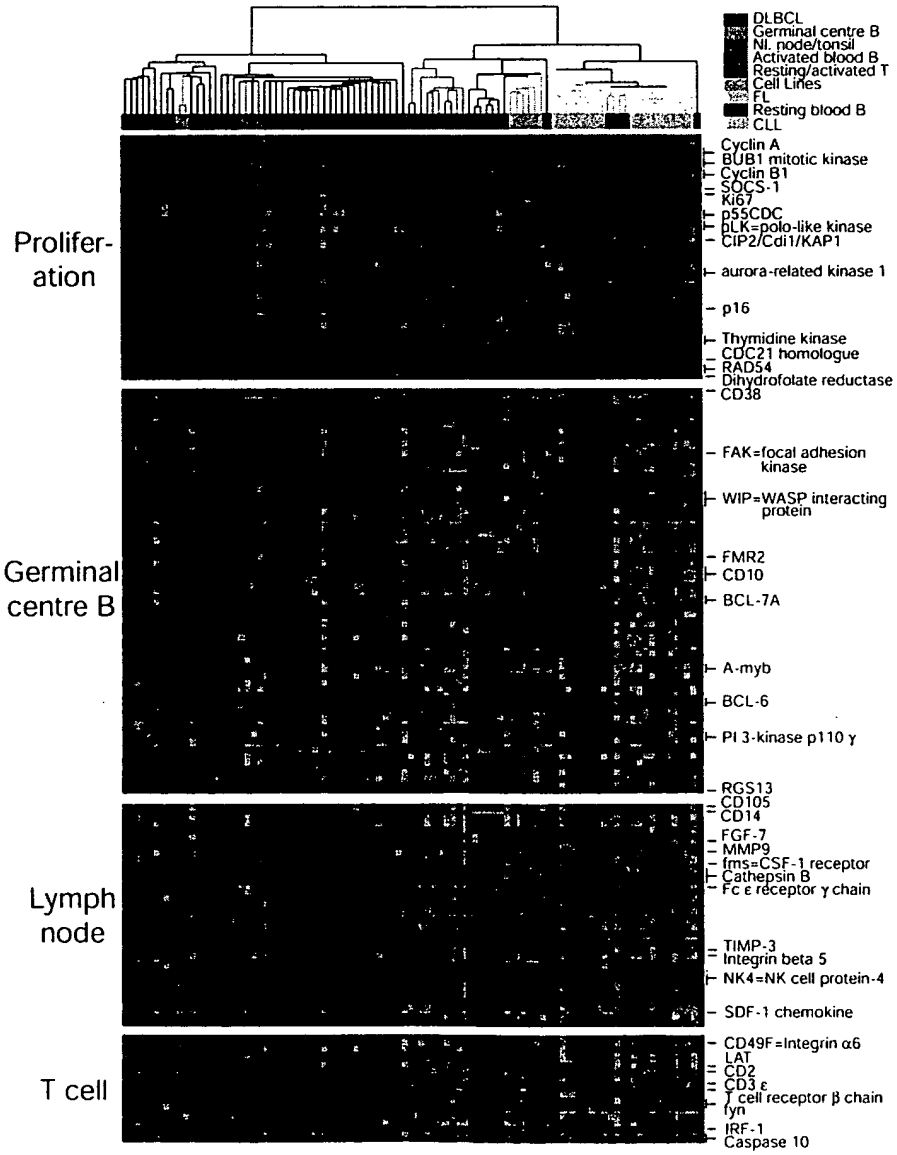


Figure 2 Expanded view of biologically distinct gene expression signatures defined by hierarchical clustering. Data are the same as in Fig. 1. Most genes without designations on the right are new genes of unknown function derived from various lymphoid cDNA libraries.

signatures could be used to define subsets of DLBCL. We decided to focus our attention initially on the germinal centre B-cell genes, however, because we suspected that these genes might identify DLBCL cases that were derived from distinct stages of normal B-cell differentiation. Indeed, the clustering of the germinal centre B-cell samples with a subset of the DLBCLs in a major branch of the dendrogram in Fig. 1 suggested that this group of DLBCLs might resemble normal germinal centre B cells.

To test this hypothesis, we reclustered the DLBCL cases using only the expression pattern of the genes that define the germinal centre B-cell signature (Fig. 3a). Two large branches were evident in the resulting dendrogram. We will refer to the groups defined by these branches as GC B-like DLBCL and activated B-like DLBCL, for reasons detailed below. The same two branches were also evident in the dendrogram in Fig. 1: activated B-like DLBCL includes all cases in the branch labelled 'A', and GC B-like DLBCL includes all cases in branch labelled 'G'. The largest DLBCL branch in Fig. 1 is a mixture of the cases assigned to the two subgroups. Normal germinal centre B cells were clustered with the GC B-like DLBCL group. Indeed, the DLBCL cases in GC B-like DLBCL group expressed, to a varying degree, all of the genes that define the germinal centre B-cell signature. In contrast, the activated B-like DLBCL group expressed these genes at low or undetectable levels, for the most part. The gene expression subgroups defined here were not obviously related to histological subtypes of DLBCL: only two of the cases studied could be assigned to the immunoblastic histological subtype, according to the revised Kiel classification system. Furthermore, no evidence of normal germinal centres was found in the lymph-node biopsies. Indeed, one of the germinal centre B-cell markers described below, CD10, was expressed by the lymphoma cells using immunohistochemistry (data not shown). These data clearly suggested that a distinct class of DLBCLs was derived from the germinal centre B cell and retained the gene expression program, and presumably many of the phenotypic characteristics, of this stage of B-cell differentiation.

We searched for genes that were selectively expressed in the activated B-like DLBCL group. This search excluded genes that were readily assigned to the proliferation, T-cell and lymph-node signatures (Fig. 1) in order to focus attention on more subtle intrinsic molecular features of this group of tumours. We used hierarchical clustering to reorder this set of 2,984 genes while maintaining the order shown in Fig. 3a of the DLBCL cases (Fig. 3b). As is evident in Fig. 3c, a cluster of genes could be recognized on the basis of their elevated expression in the activated B-like DLBCLs, as compared with GC B-like DLBCLs. It is important to note that considerable gene expression heterogeneity exists within each subgroup and that no single gene in either of these large clusters was absolutely correlated in expression with the DLBCL subgroup taxonomy. Rather, patients assigned by this method to either DLBCL subgroup shared a large gene expression program that distinguished them from the other subgroup.

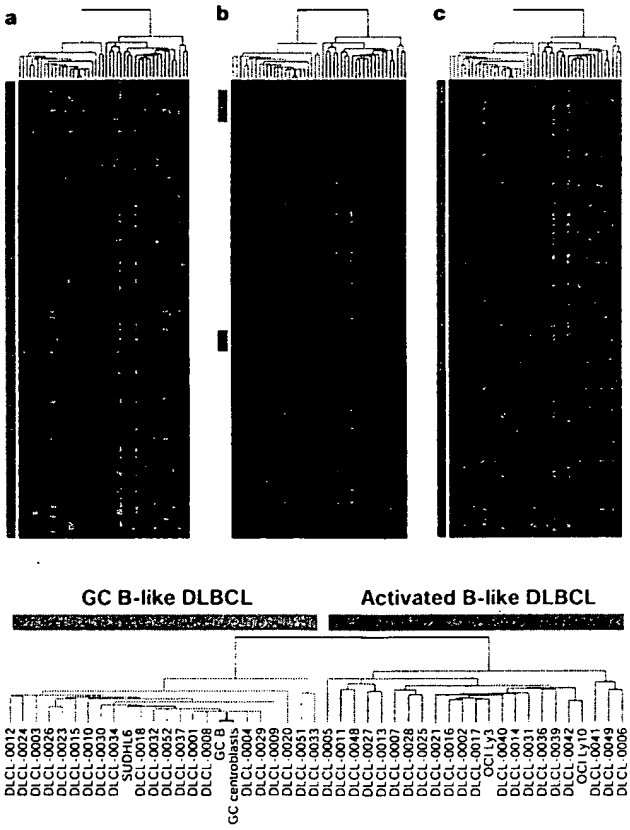


Figure 3 Discovery of DLBCL subtypes by gene expression profiling. The samples used in this clustering analysis are shown at the bottom. **a**, Hierarchical clustering of DLBCL cases (blue and orange) and germinal centre B cells (black) based on the genes of the germinal centre B-cell gene expression signature shown in Figs 1 and 2. Two DLBCL subgroups, GC B-like DLBCL (orange) and activated B-like DLBCL (blue) were defined by this process. **b**, Discovery of genes that are selectively expressed in GC B-like DLBCL and activated B-like DLBCL. All genes from Fig. 1, with the exception of the genes in the proliferation, T-cell and lymph-node gene expression signatures, were ordered by hierarchical clustering while maintaining the order of samples determined in Fig. 3a. Genes selectively expressed in GC B-like DLBCL (orange) and activated B-like DLBCL (blue) are indicated. **c**, Hierarchical clustering of the genes selectively expressed in GC B-like DLBCL and activated B-like DLBCL, which was determined from Fig. 3b.

DLBCL subgroups and B-cell differentiation

We examined how all of the genes that distinguish these DLBCL subgroups are expressed during B-cell differentiation and activation. Figure 4 shows that almost all of the genes that defined GC B-like DLBCL were highly expressed in normal germinal centre B cells. Most of these genes were expressed at low or undetectable levels in peripheral blood B cells that had been activated *in vitro* by a variety of mitogenic signals. Some of the GC B-like DLBCL genes were expressed in resting blood B cells and germinal centre B cells at comparable levels but not in activated peripheral blood B cells. Conversely, virtually all of the genes that were selectively expressed in germinal centre B cells relative to resting or activated peripheral blood B cells were expressed by GC B-like DLBCL (data not shown).

By contrast, most of the genes that defined activated B-like DLBCL were not expressed in normal germinal centre B cells (Fig. 4). Instead, many of these genes, but not all, were induced during *in vitro* activation of peripheral blood B cells. The time course of expression of these genes during B-cell activation varied, with some genes induced after 6 h of activation and others only expressed after 48 h of activation. Thus, the gene expression signature of activated B-like DLBCLs is reminiscent of, but not identical to, the signature of activated peripheral blood B cells. Notably, two DLBCL cell lines, OCI Ly3 and OCI Ly10, were among the activated B-like DLBCLs. In fact, one or both of these two cell lines expressed virtually all of the genes that defined the activated B-like DLBCL signature. This observation suggests that signal transduction pathways that are inducibly engaged during peripheral B-cell activation and mitogenesis are constitutively active in activated B-like DLBCLs.

The gene expression program that distinguishes GC B-like DLBCLs includes many known markers of germinal centre differentiation (for example, the genes encoding the cell-surface proteins CD10 and CD38 (ref. 24), the nuclear factor A-myb (ref. 25) and the DNA repair protein 8-oxoguanine DNA glycosylase (OGG1)[26]) and a host of new genes. A particularly noteworthy gene in the GC B-like DLBCL signature is BCL-6, a well-established

germinal centre marker that is also the most frequently translocated gene in DLBCL[22]. Although BCL-6 protein expression is invariably detected in DLBCL, its levels vary and are not correlated with the presence of BCL-6 translocations[27,28]. Cytogenetic data are available for 16 out of the DLBCL cases studied here and do not support a link between elevated BCL-6 mRNA levels in GC B-like DLBCL and BCL-6 translocations (data not shown). Thus, the higher expression of BCL-6 mRNA in GC B-like DLBCLs is most probably related to their derivation from germinal centre B cells (Fig. 4).

Two other genes that can be altered by translocations in lymphoid malignancies, BCL-7A and LMO2 (TTG-2/RBTN2), have not previously been described as highly expressed in germinal centre B cells. BCL-7A was cloned as part of a complex chromosomal translocation in a Burkitt's lymphoma cell line and was found to be rearranged in another cell line derived from mediastinal large B-cell lymphoma[29]. The specific expression of BCL-7A in germinal centre B cells has strong parallels with BCL-6. BCL-6 is required for germinal centre formation during an antigen-driven immune response[30–32] and is translocated in B-cell malignancies that derive from germinal centre B cells. Given the preferential expression of BCL-7A in germinal centre B cells, it is conceivable that this gene is also involved in normal germinal centre physiology and in the pathophysiology of GC B-like DLBCL. LMO2 is translocated and overexpressed in a subset of T-cell acute lymphoblastic leukaemias

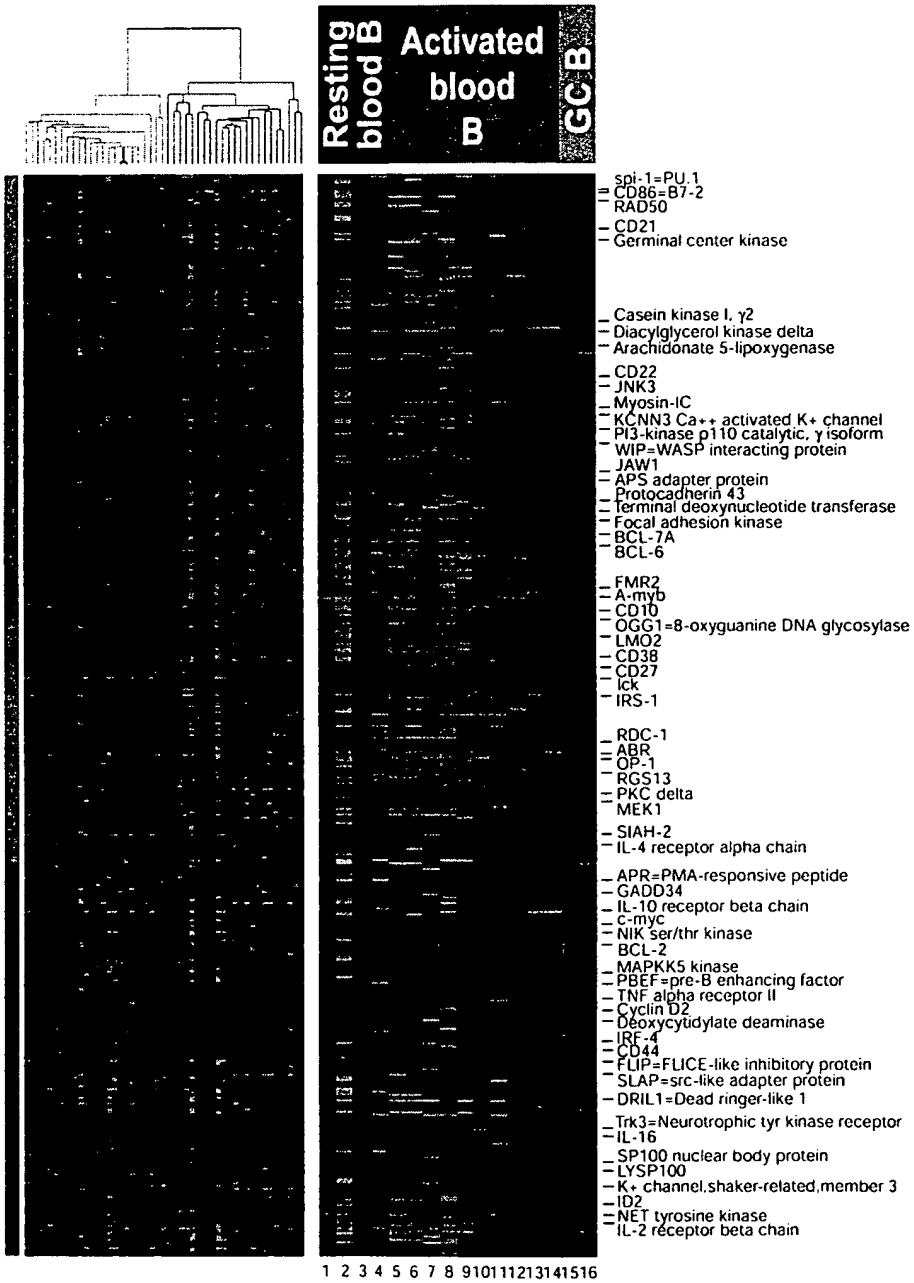


Figure 4 Relationship of DLBCL subgroups to normal B-lymphocyte differentiation and activation. The data in the left panel are taken from Fig. 3c. The right panel depicts gene expression data from the following normal B-cell samples: (1) Total CD19$^+$ blood B cells; (2) Naive CD27$^-$ blood B cells; (3) Memory CD27$^+$ blood B cells; (4) cord blood CD19$^+$ B cells; (5) blood B cells; anti-IgM 6 h; (6) blood B cells; anti-IgM + IL-4 6 h; (7) blood B cells; anti-IgM + CD40 ligand 6 h; (8) blood B cells; anti-IgM + CD40 ligand + IL-4 6 h; (9) blood B cells; anti-IgM 24 h; (10) blood B cells; anti-IgM + IL-4 24 h; (11) blood B cells; anti-IgM + CD40 ligand 24 h; (12) blood B cells; anti-IgM + CD40 ligand + IL-4 24 h; (13) blood B cells; anti-IgM + CD40 ligand (low concentration) 48 h; (14) blood B cells; anti-IgM + CD40 ligand (high concentration) 48 h; (15) tonsil germinal centre B cells; (16) tonsil germinal centre centroblasts. See Supplementary Information for full data.

and LMO2 transgenic mice have a block in early T-cell differentiation and develop T-cell leukaemia[33]. The selective expression of LMO2 in germinal centre B cells indicates that LMO2 may have a role in inhibiting differentiation in the B-cell lineage as well, and perhaps a corresponding role in the DLBCL malignant phenotype.

The activated B-like DLBCL signature also includes a gene that is translocated in lymphoid malignancies, IRF4 (MUM1/LSIRF). IRF4 is fused to the immunoglobulin locus in some cases of multiple myeloma and can function as an oncogene *in vitro*[34]. IRF4 is transiently induced during normal lymphocyte activation[35] (Fig. 4) and is critical for the proliferation of B lymphocytes in response to signals from the antigen receptor[36]. Thus, the constitutive expression of IRF4 in activated B-like DLBCLs may contribute to the unchecked proliferation of the malignant cells in these tumours.

A notable feature of the gene expression pattern of activated B-like DLBCLs was the expression of two genes whose products inhibit programmed cell death. FLIP (FLICE-like inhibitory protein/I-FLICE/FLAME-1/Casper/MRIT/CASH/CLARP) is a dominant-negative mimic of caspase 8 (FLICE) which can block apoptosis mediated by Fas and other death receptors[37]. FLIP is induced early during normal lymphocyte activation, presumably to block activation-induced apoptosis that occurs physiologically later in an immune response. FLIP is highly expressed in many tumour types, and its constitutive expression in activated B-like DLBCLs could inhibit apoptosis of tumour cells induced by host T cells expressing Fas ligand[38,39]. The key anti-apoptotic gene BCL-2 is translocated in most cases of follicular lymphoma and in a subset of DLBCL. BCL-2 mRNA is not expressed in germinal centre B cells but is induced more than 30-fold during activation of peripheral blood B cells (Fig. 4). Most activated B-like DLBCLs (71%) had BCL-2 mRNA levels more than fourfold higher than were observed in germinal centre B cells (Fig. 4). This overexpression did not correlate with BCL-2 translocations (data not shown). A minority of GC B-like DLBCLs (29%) had similarly elevated BCL-2 mRNA levels, indicating that BCL-2 may also be important in some cases of this DLBCL subgroup.

DLBCL gene expression subgroups define prognostic categories

Does the taxonomy of DLBCL derived from gene expression patterns define clinically distinct subgroups of patients? None of the patients included in this study had been treated before obtaining the biopsy sample. Furthermore, these patients were '*de novo*' DLBCL cases that had not obviously arisen from pre-existing low-grade malignancies such as follicular lymphoma. After biopsy, the patients were treated at two medical centres using comparable, standard multi-agent chemotherapy regimens. Figure 5a presents a Kaplan–Meier plot of overall survival data from these patients, segregated according to gene expression subgroup. Germinal centre B-like and activated B-like DLBCLs were associated with statistically significant differences in overall survival ($P < 0.01$) and in event-free survival (data not shown). Although the average five-year survival for all patients was 52%, 76% of GC B-like DLBCL patients were still alive after five years, as compared with only 16% of activated B-like DLBCL patients. The differential survival of patients in the two DLBCL subgroups was apparently uninfluenced by the anthracyline-based chemotherapeutic regimen used (data not shown), which is not surprising as responses of DLBCL patients to various multi-agent chemotherapeutic regimens were found to be equivalent[40]. Thus, the molecular differences between these two kinds of lymphoma were accompanied by a remarkable divergence in clinical behaviour, suggesting that GC B-like DLBCL and activated B cell DLBCL should be regarded as distinct diseases.

A clinical indicator of prognosis, the International Prognostic Indicator (IPI), has been successfully used to define prognostic subgroups in DLBCL[8]. This indicator takes into account the patient's age, performance status, and the extent and location of disease. As suspected, within our patient population a low IPI score (0–2) identified patients with better overall survival as compared with patients with a high IPI score (3–5) (Fig. 5b). We then determined whether our molecular definition of DLBCL subgroups could add to the prognostic value of this clinical indicator of prognosis. Considering only patients with low clinical risk, as judged by the IPI, patients in the activated B-like DLBCL group had a distinctly worse overall survival than patients in the GC B-like DLBCL group ($P < 0.05$) (Fig. 5c). Thus, the molecular dissection of DLBCL by gene expression profiling and the IPI apparently identify different features of these patients that influence their survival.

Conclusions

This study shows that a genomic view of gene expression in cancer can bring clarity to previously muddy diagnostic categories. The precision of morphological diagnosis, even when supplemented with immunohistochemistry for a few markers, was insufficient in the case of DLBCL to identify believable diagnostic subgroups. A number of individual markers have been used to define subsets of DLBCL[41–46], but these studies do not provide the present overview that strongly implies that this single diagnostic category of lymphoma harbours at least two distinct diseases. Indeed, the new methods of gene expression profiling call for a revised definition of what is deemed a 'disease'. The two DLBCL subgroups are distinguished from each other by the differential expression of hundreds of different genes, and these genes relate each subgroup to a separate stage of B-cell differentiation and activation. These molecular differences, in the light of accompanying clinical

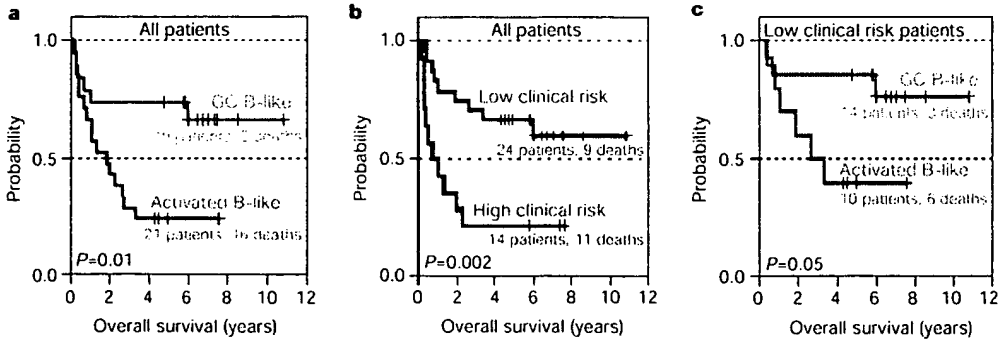


Figure 5 Clinically distinct DLBCL subgroups defined by gene expression profiling. **a**, Kaplan–Meier plot of overall survival of DLBCL patients grouped on the basis of gene expression profiling. **b**, Kaplan–Meier plot of overall survival of DLBCL patients grouped according to the International Prognostic Index (IPI). Low clinical risk patients (IPI score 0–2) and high clinical risk patients (IPI score 3–5) are plotted separately. **c**, Kaplan–Meier plot of overall survival of low clinical risk DLBCL patients (IPI score 0–2) grouped on the basis of their gene expression profiles.

differences between these subgroups, suggest that these two subgroups of DLBCL should be considered separate diseases.

Nonetheless, we do not wish to imply that patients within a DLBCL subgroup defined here are monomorphic. As mentioned above, considerable molecular heterogeneity exists within each DLBCL subgroup. As many more DLBCL patients are studied by gene expression profiling, it is quite possible that more subgroups will emerge. Given that many current diagnostic categories of non-Hodgkin's lymphoma constitute less than 10% of the total cases[5], it seems likely that the DLBCL diagnostic category will also include a number of minor subgroups.

The classification scheme highlighted in this study divided DLBCL on the basis of genes that are differentially expressed within the B-cell lineage. This particular classification identified patient groups that differed in survival after treatment with anthracycline-based multi-agent chemotherapy regimens. It is unclear at present which of the genes that distinguish GC B-like DLBCL from activated B-like DLBCL are the most important molecular determinants of chemotherapy responsiveness. Furthermore, there is residual clinical heterogeneity which cannot be explained by the current classification. Despite the fact that patients with GC B-like DLBCL had an overall favourable prognosis, five patients died within the first two years of diagnosis. Likewise, three patients in the activated B-like DLBCL subgroup were alive five years after treatment, despite the poor outcome of most patients in this subgroup. By profiling the gene expression of many more DLBCLs, it may become possible to implicate a single gene or pathway in chemotherapy responsiveness with statistical certainty. More probably, however, a multivariate approach to prognosis will be needed that combines knowledge of the DLBCL subgroup, as defined here, with measurements of individual genes or pathways that contribute to treatment outcome.

Gene expression profiling presents a new way of approaching cancer therapeutics in the future. Current treatment of DLBCL typically begins with multi-agent chemotherapy, and then, if a complete remission cannot be maintained, patients are considered for bone marrow transplantation[7]. The definition of prognostic groups by gene expression profiling, in combination with clinical indicators such as the IPI, may lead to the recommendation that some patients receive early bone marrow transplantations upon initial diagnosis. In testing cancer therapeutics in clinical trials, it is obviously beneficial to define homogeneous populations of patients to improve the likelihood of observing efficacy in specific disease entities. We anticipate that global surveys of gene expression in cancer, such as we present here, will identify a small number of marker genes that will be used to stratify patients into molecularly relevant categories which will improve the precision and power of clinical trials.

Finally, the genomic-scale view of gene expression in cancer provides a unique perspective on the development of new cancer therapeutics that could be based on a molecular understanding of the cancer phenotype. Our study shows that the two DLBCL subgroups differentially expressed entire transcriptional modules composed of hundreds of genes, many of which could be expected to contribute to the malignant behaviour of the tumour. This observation suggests that successful new therapeutics might be aimed at the upstream signal-transducing molecules whose constitutive activity in these lymphomas leads to expression of pathological transcriptional programs. □

Methods

Messenger RNA samples

Total germinal centre B cells and centroblasts were purified from human tonsils as described[24]. Human blood B cells were purified from adult apheresis products or cord blood by magnetic enrichment for CD19$^+$ cells (Miltenyi Biotec). Naive CD27$^+$ B cells and memory CD27$^-$ blood B cells were isolated by fluorescent cell sorting starting with CD19$^+$ adult peripheral blood B cells[47,48]. Magnetic cell sorting was used to purify CD4$^+$, CD45RA^{high} T cells from human cord blood or adult peripheral blood and CD4$^+$ thymocytes from human fetal thymus (Milteni Biotec). All lymphocyte samples were purified to more than 98% homogeneity as determined by FACS analysis. For rare lymphoid subpopulations such as centroblasts or resting and naive peripheral blood B cells, purified samples from multiple donors were pooled for microarray analysis. *In vitro* stimulation of peripheral B cells was done as described[49] using anti-IgM antibody, IL-4 and/or CD40 ligand-containing membranes. Most experiments used a 1:1000 dilution of CD40 ligand membranes (designated 'low' concentration, Figs 1 and 4) but one experiment used a 1:200 dilution (designated 'high' concentration, Figs 1 and 4). T cells were stimulated for 2 h with phorbol ester (50 ng ml^{-1}) and ionomycin (1.5 μM). Patient samples were obtained after informed consent and were treated anonymously during microarray analysis. DLBCL patients were treated at either University of Nebraska Medical Center (n = 34) or Stanford University School of Medicine (n = 8) using comparable, anthracycline-based, multi-agent chemotherapeutic regimens with curative intent. Clinical data were not available on two of the DLBCL cases presented in Fig. 1 (DLCL-51 and DLCL-52). For two additional patients (DLCL-25 and DLCL-36), the data needed to calculate the IPI were not available. DLBCL and FL lymph-node biopsies were either snap frozen, frozen in OCT or disaggregated and frozen as a viable cell suspension. Chronic lymphocyte leukaemia cells were purified from untreated patients by magnetic selection for CD19$^+$ cells (Miltenyi Biotec).

Microarray procedures

DNA microarray analysis of gene expression was done essentially as described[50]. The cDNA clones on the Lymphochip microarray are listed in Supplementary Information and are available from Research Genetics. Fluorescent images of hybridized microarrays were obtained using a GenePix 4000 microarray scanner (Axon Instruments). Images were analysed with ScanAlyze (M. Eisen; http://www.microarrays.org/software), and fluorescence ratios (along with numerous quality control parameters; see ScanAlyze manual) were stored in a custom database. Single spots or areas of the array with obvious blemishes were flagged and excluded from subsequent analyses. Raw data files for each array containing all measured values and manual flags are available in Supplementary Information. A set of clones that consistently behaved poorly across arrays was identified and excluded from all analyses (see Supplementary Information). Fluorescence ratios were calibrated independently for each array by applying a single scaling factor to all fluorescent ratios from each array; this scaling factor was computed so that the median fluorescence ratio of well-measured spots on each array was 1.0.

All cDNA microarray analyses were performed using poly-(A)$^+$ mRNA (Fast Track, Invitrogen). In each experiment, fluorescent cDNA probes were prepared from an experimental mRNA sample (Cy5-labelled) and a control mRNA sample (Cy3-labelled) isolated from a pool of nine lymphoma cell lines (Raji, Jurkat, L428, OCI-Ly3, OCI-Ly8, OCI-Ly1, SUDHL5, SUDHL6 and WSU1). The use of a common control cDNA probe allows the relative expression of each gene to be compared across all samples[18].

Data analysis

All non-flagged array elements for which the fluorescent intensity in each channel was greater than 1.4 times the local background were considered well measured. The ratio values were log-transformed (base 2) and stored in a table (rows, individual cDNA clones; columns, single mRNA samples). Where samples had been analysed on multiple arrays, multiple observations for an array element for a single sample were averaged. Array elements that were not well measured on at least 80% of the 96 mRNA samples were excluded. Data for the remaining genes were centred by subtracting (in log space) the median observed value, to remove any effect of the amount of RNA in the reference pool. This dataset contains 4,026 array elements (see Supplementary Information). Hierarchical clustering was applied to both axes using the weighted pair-group method with centroid average as implemented in the program Cluster (M. Eisen; http://www.microarrays.org/software)[20]. The distance matrixes used were Pearson correlation for clustering the arrays and the inner product of vectors normalized to magnitude 1 for the genes (this is a slight variant of Pearson correlation; see Cluster manual available at http://www.microarrays.org/software/ for computational details). The results were analysed with Tree View (M. Eisen; http://www.microarrays.org/software)[20]. All datasets and image files used to generate Figs 1–4 are included in the Supplementary Information, along with numerous supplementary and additional analyses.

Received 11 November 1999; accepted 10 January 2000.

1. Hodgkin, T. On some morbid appearances of the absorbant glands and spleen. *Med.-Chir. Trans.* **17**, 68–114 (1832).
2. Sternberg, C. Über eine eigenartige unter dem Bilde der Pseudoleukamie verlaufende Tuberculose des lymphatischen Apparates. *Heilk* **19**, 21–90 (1898).
3. Reed, D. M. On the pathological changes in Hodgkin's disease, with especial reference to its relation to tuberculosis. *Johns Hopkins Hosp. Rep.* **10**, 133–196 (1902).
4. Rosenberg, S. A. Classification of lymphoid neoplasms. *Blood* **84**, 1359–1360 (1994).
5. Harris, N. L. *et al.* A revised European–American classification of lymphoid neoplasms: a proposal from the International Lymphoma Study Group. *Blood* **84**, 1361–1392 (1994).
6. The Non-Hodgkin's Lymphoma Classification Project: A clinical evaluation of the International Lymphoma Study Group classification of non-Hodgkin's lymphoma. *Blood* **89**, 3909–3918 (1997).
7. Vose, J. M. Current approaches to the management of non-Hodgkin's lymphoma. *Semin. Oncol.* **25**, 483–491 (1998).
8. The International Non-Hodgkin's Lymphoma Prognostic Factors Project: A predictive model for aggressive non-Hodgkin's lymphoma. *N. Engl. J. Med.* **329**, 987–994 (1993).
9. Klein, U. *et al.* Somatic hypermutation in normal and transformed human B cells. *Immunol. Rev.* **162**, 261–280 (1998).

10. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
11. Bubendorf, L. *et al.* Hormone therapy failure in human prostate cancer: analysis by complementary DNA and tissue microarrays. *J. Natl Cancer Inst.* **91**, 1758–1764 (1999).
12. Wang, K. *et al.* Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* **229**, 101–108 (1999).
13. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
14. Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
15. Perou, C. M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96**, 9212–217 (1999).
16. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* **14**, 457–460 (1996).
17. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
18. Alizadeh, A. *et al.* The Lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harbor Symp. Quant. Biol.* (in the press).
19. Alizadeh, A., Eisen, M., Botstein, D., Brown, P. O. & Staudt, L. M. Probing lymphocyte biology by genomic-scale gene expression analysis. *J. Clin. Immunol.* **18**, 373–379 (1998).
20. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
21. Grogan, T. M. *et al.* Independent prognostic significance of a nuclear proliferation antigen in diffuse large cell lymphomas as determined by the monoclonal antibody Ki-67. *Blood* **71**, 1157–1160 (1988).
22. Staudt, L. M., Dent, A. L., Shaffer, A. L. & Yu, X. Regulation of lymphocyte cell fate decisions and lymphomagenesis by BCL-6. *Int. J. Immunol.* **18**, 381–403 (1999).
23. Bahler, D. W. & Levy, R. Clonal evolution of a follicular lymphoma: evidence for antigen selection. *Proc. Natl Acad. Sci. USA* **89**, 6770–6774 (1992).
24. Liu, Y. -J. & Banchereau, J. in *Handbook of Experimental Immunology* (eds Weir, D., Blackwell, C., Herzenberg, L. & Herzenberg, L.) 93.1–93.9 (Blackwell Scientific, Oxford, 1996).
25. Golay, J., Erba, E., Bernasconi, S., Peri, G. & Introna, M. The A-myb gene is preferentially expressed in tonsillar CD38+, CD39–, and sIgM– B lymphocytes and in Burkitt's lymphoma cell lines. *J. Immunol.* **153**, 543–553 (1994).
26. Kuo, F. C. & Sklar, J. Augmented expression of a human gene for 8-oxoguanine DNA glycosylase (MutM) in B lymphocytes of the dark zone in lymph node germinal centers. *J. Exp. Med.* **186**, 1547–1556 (1997).
27. Flenghi, L. *et al.* A specific monoclonal antibody (PG-B6) detects expression of the BCL-6 protein in germinal center B cells. *Am. J. Pathol.* **147**, 405–411 (1995).
28. Pittaluga, S. *et al.* BCL-6 expression in reactive lymphoid tissue and in B-cell non-Hodgkin's lymphomas. *J. Pathol.* **179**, 145–150 (1996).
29. Zani, V. J. *et al.* Molecular cloning of complex chromosomal translocation t(8;14;12)(q24. 1;q32. 3;q24. 1) in a Burkitt lymphoma cell line defines a new gene (BCL7A) with homology to caldesmon. *Blood* **87**, 3124–3134 (1996).
30. Fukuda, T. *et al.* Disruption of the Bcl6 gene results in an impaired germinal center formation. *J. Exp. Med.* **186**, 439–448 (1997).
31. Ye, B. H. *et al.* The BCL-6 proto-oncogene controls germinal-centre formation and Th2-type inflammation. *Nature Genet.* **16**, 161–170 (1997).
32. Dent, A. L., Shaffer, A. L., Yu, X., Allman, D. & Staudt, L. M. Control of inflammation, cytokine expression, and germinal center formation by BCL-6. *Science* **276**, 589–592 (1997).
33. Rabbitts, T. H. LMO T-cell translocation oncogenes typify genes activated by chromosomal translocations that alter transcription and developmental processes. *Genes Dev.* **12**, 2651–2657 (1998).
34. Iida, S. *et al.* Deregulation of MUM1/IRF4 by chromosomal translocation in multiple myeloma. *Nature Genet.* **17**, 226–230 (1997).
35. Matsuyama, T. *et al.* Molecular cloning of LSIRF, a lymphoid-specific member of the interferon regulatory factor family that binds the interferon-stimulated response element (ISRE). *Nucleic Acids Res.* **23**, 2127–2136 (1995).
36. Mittrucker, H. W. *et al.* Requirement for the transcription factor LSIRF/IRF4 for mature B and T lymphocyte function. *Science* **275**, 540–543 (1997).
37. Tschopp, J., Irmler, M. & Thome, M. Inhibition of fas death signals by FLIPs. *Curr. Opin. Immunol.* **10**, 552–558 (1998).
38. Djerbi, M. *et al.* The inhibitor of death receptor signaling, FLICE-inhibitory protein defines a new class of tumor progression factors. *J. Exp. Med.* **190**, 1025–1031 (1999).
39. Medema, J. P., de Jong, J., van Hall, T., Melief, C. J. M. & Offringa, R. Immune escape of tumors *in vivo* by expression of cellular FLICE-inhibitory protein. *J. Exp. Med.* **190**, 1033–1038 (1999).
40. Fisher, R. I. *et al.* Comparison of a standard regimen (CHOP) with three intensive chemotherapy regimens for advanced non-Hodgkin's lymphoma. *N. Engl. J. Med.* **328**, 1002–1006 (1993).
41. Jalkanen, S., Joensuu, H., Soderstrom, K. O. & Klemi, P. Lymphocyte homing and clinical behavior of non-Hodgkin's lymphoma. *J. Clin. Invest.* **87**, 1835–1840 (1991).
42. Harada, S. *et al.* Molecular and immunological dissection of diffuse large B cell lymphoma: CD5–, and CD5– with CD10+ groups may constitute clinically relevant subtypes. *Leukemia* **13**, 1441–1447 (1999).
43. Kramer, M. H. *et al.* Clinical significance of bcl2 and p53 protein expression in diffuse large B-cell lymphoma: a population-based study. *J. Clin. Oncol.* **14**, 2131–2138 (1996).
44. Preti, H. A. *et al.* Prognostic value of serum interleukin-6 in diffuse large-cell lymphoma. *Ann. Int. Med.* **127**, 186–194 (1997).
45. Gascoyne, R. D. *et al.* Prognostic significance of Bcl-2 protein expression and Bcl-2 gene rearrangement in diffuse aggressive non-Hodgkin's lymphoma. *Blood* **90**, 244–251 (1997).
46. Kramer, M. H. *et al.* Clinical relevance of BCL2, BCL6, and MYC rearrangements in diffuse large B-cell lymphoma. *Blood* **92**, 3152–3162 (1998).
47. Klein, U., Rajewsky, K. & Kuppers, R. Human immunoglobulin (Ig)M+IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. *J. Exp. Med.* **188**, 1679–1689 (1998).
48. Tangye, S. G., Liu, Y. J., Aversa, G., Phillips, J. H. & de Vries, J. E. Identification of functional human splenic memory B cells by expression of CD148 and CD27. *J. Exp. Med.* **188**, 1691–1703 (1998).
49. Allman, D. *et al.* BCL-6 expression during B-cell activation. *Blood* **87**, 5257–5268 (1996).
50. Eisen, M. B. & Brown, P. O. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303**, 179–205 (1999).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com), on the authors' World-Wide Web site (http://llmpp.nih.gov/lymphoma) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We acknowledge the support of the Cancer Genome Anatomy Project (CGAP), led by B. Strausberg and R. Klausner. We also thank R. Klausner for comments on the manuscript; C. Prange for providing CGAP cDNA clones; H. Messner for providing DLBCL cell lines; H. Mostowski for sorting lymphocyte subpopulations by FACS; Holy Cross Hospital, Silver Spring, Maryland, for providing human tonsils; J. DeRisi for helpful advice on microarray technology; and members of the Staudt, Brown and Botstein laboratories for helpful discussions. Research at Stanford was supported by grants from the National Cancer Institute to D.B., R.L. and P.O.B. and by the Howard Hughes Medical Institute. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute. A.A. was initially supported by the Howard Hughes Medical Institute Research Scholar Program while at the NIH and then by the Medical Scientist Training Program at Stanford University. M.B.E. was supported by a Computational Molecular Biology Postdoctoral Fellowship from the Alfred E. Sloan Foundation.

Correspondence and requests for materials should be addressed to L.M.S. (lstaudt@box-l.nih.gov) or P.O.B. (pbrown@cmgm.stanford.edu).

RELATED PROCEEDINGS APPENDIX

There are no other appeals or interferences related to the instant appeal.